HSC Research Report

HSC/13/07

An empirical comparison of alternate schemes for combining electricity spot price forecasts

Jakub Nowotarski*
Eran Raviv**
Stefan Trueck***
Rafał Weron*

* Institute of Organization and Management, Wrocław University of Technology, Poland
** Department of Econometrics, Erasmus University, Rotterdam, The Netherlands
*** Faculty of Business and Economics, Macquarie University, Sydney, Australia

Hugo Steinhaus Center
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
http://www.im.pwr.wroc.pl/~hugo/

An empirical comparison of alternate schemes for combining electricity spot price forecasts

Jakub Nowotarski[a], Eran Raviv[b], Stefan Trück[c], Rafał Weron[a]

[a]*Institute of Organization and Management, Wrocław University of Technology, Wrocław, Poland*
[b]*Department of Econometrics, Erasmus University, Rotterdam, The Netherlands*
[c]*Faculty of Business and Economics, Macquarie University, Sydney, Australia*

Abstract

In this paper we investigate the use of forecast averaging for electricity spot prices. While there is an increasing body of literature on the use of forecast combinations, there is only a small number of applications of these techniques in the area of electricity markets. In this comprehensive empirical study we apply seven averaging and one selection scheme and perform a backtesting analysis on day-ahead electricity prices in three major European and US markets. Our findings support the additional benefit of combining forecasts for deriving more accurate predictions, however, the performance is not uniform across the considered markets. Interestingly, equally weighted pooling of forecasts emerges as a viable robust alternative compared with other schemes that rely on estimated combination weights. Overall, we provide empirical evidence that also for the extremely volatile electricity markets, it is beneficial to combine forecasts from various models for the prediction of day-ahead electricity prices. In addition, we empirically demonstrate that not all forecast combination schemes are recommended.

Keywords: Electricity price forecasting, Forecasts combination, ARX model, Day-ahead market

1. Introduction

Since the early 1990s, structural reforms and deregulation have lead to significant changes in worldwide electricity markets. Like other commodities, electricity is now traded under competitive rules using spot and derivative contracts (Bunn, 2004; Shahidehpour et al., 2002). However, one particular feature of most electricity markets is that the spot market is actually a day-ahead market that does not allow for continuous trading. This is a result of system operators requiring advance notice in order to verify that the schedule is feasible and lies within transmission constraints. In a day-ahead market agents submit their bids and offers for delivery of electricity during each hour (or half-hour) of the next day before a certain market closing time. Thus, when dealing with the modeling and forecasting of intraday prices it is important to recall that prices for all spot contracts of the next day are determined at the same time using the same available information (Conejo et al., 2005; Huisman et al., 2007; Misiorek et al., 2006; Peña, 2012). The system price is then calculated as the equilibrium point for the aggregated supply and demand curves and for each of the hourly or half-hourly intervals.

Email addresses: jakub.nowotarski@gmail.com (Jakub Nowotarski), raviv@ese.eur.nl (Eran Raviv), stefan.trueck@mq.edu.au (Stefan Trück), rafal.weron@pwr.wroc.pl (Rafał Weron)

August 21, 2013

In contrast to other tradable commodities, electricity is practically non-storable. As a result the electricity spot (or day-ahead) price time series exhibit specific characteristics. The seasonal character of the prices is a direct consequence of demand fluctuations that mostly arise from business hours at the daily or weekly level or changing climate conditions like temperature or the number of daylight hours at the yearly scale. In addition to seasonality and mean reversion, electricity prices exhibit an extremely high price volatility as well as infrequent, but large price spikes. These features have forced producers and wholesale consumers to hedge not only against volume risk but also against price movements. This in turn has significantly enhanced research efforts towards modeling and forecasting spot electricity prices.

A wide range of econometric or statistical models have been suggested in the literature including regression models, jump-diffusions, GARCH-type models and regime-switching models (for reviews see e.g. Eydeland and Wolyniec, 2013; Huisman, 2009; Weron, 2006). However, in terms of predicting spot price movements, each model specification yields a different forecast. Facing the variety of alternative models available in the literature, one could discard all of the models but one on the basis of their goodness of fit and forecasting performance. Alternatively, one can allocate weights to the various forecasts produced by individual models in order to obtain a combined forecast for the day-ahead electricity price. The latter strategy may be more favorable in the context of changing model and predictor relevance through time and can potentially achieve a better forecasting performance by virtue of smoothed model selection. The best model is not known in advance so allocating weights to the individual models is used as a hedge against the possibility of an inaccurate model choice. That said, this strategy can also decrease the overall accuracy, if the best model is easy to recognize beforehand.

Despite the increasing body of literature on the use of forecast combinations (also referred to as 'combining forecasts', 'forecast averaging' or 'model averaging') for prediction, there is only a small number of applications of these techniques in the area of electricity markets. To our best knowledge, existing applications so far only include the work by Løland et al. (2012), Smith (1989), Taylor (2010) and Taylor and Majithia (2000) in the context of load and transmission congestion forecasting, and by Bordignon et al. (2013) and Raviv et al. (2013) in the context of spot (day-ahead) price forecasting. The relatively small number of studies on combining forecasts produced by various models is surprising since, in the context of electricity price forecasting, research shows that performance of individual models is often unstable and dependent on the considered periods of price behavior, see e.g. Conejo et al. (2005) and Weron and Misiorek (2008). This motivates us to thoroughly investigate whether forecast combinations are able to outperform individual methods.

The contribution of our paper is twofold. First, we apply a great variety of stochastic models and forecast combination techniques to the data. These include, for example, standard autoregressive models, regime-switching models, mean-reversion jump diffusion models and semi-parametric autoregressive models. Techniques for forecast combinations include simple equal weighted averaging, forecast combinations based on OLS regression, constrained least squares regression (CLS, PW), Least Absolute Deviation (LAD) regression, as well as model averaging based on a Bayesian approach. The majority of the averaging techniques have not been applied to forecasting electricity spot prices yet. Second, we provide the so far most extensive study, using four datasets from key electricity markets worldwide. Considered markets include the Nordic power exchange (Nord Pool), the European Energy Exchange in Leipzig (EEX) and the Pennsylvania-New Jersey-Maryland Interconnection (PJM). For these markets we compare the averaging techniques with the realistic situation where the market participants have to decide ex ante which individual model to use. Hereby, we assume that participants decide to pick one

of the models that performed well in the past, and then examine the performance of this model in comparison to the averaging techniques. We evaluate the performance based on different criteria and conduct Diebold-Mariano tests in order to investigate whether combining forecasts can significantly improve the performance. Our findings suggest that several of the examined averaging schemes do generally provide better results than using individual models only.

The remainder of the article is organized as follows. Section 2 provides an overview of the recent literature on forecast averaging and its limited applications in electricity markets. Section 3 describes the four datasets used in this study, while Section 4 reviews the individual models for forecasting electricity spot prices and the averaging techniques. Finally, Section 5 presents empirical results and Section 6 concludes.

2. Combining forecasts and electricity markets

The idea of combining forecasts goes back to the late 1960s, with the works of Bates and Granger (1969), Crane and Crotty (1967) and Newbold and Granger (1974). Examining forecast combinations using various models and weights based on mean squared errors, the authors found a significant improvement in terms of reducing prediction errors. Since then, many authors have suggested the superior performance of forecast combinations over the use of individual models, see e.g. Clemen (1989), Diebold and Pauly (1987), de Menezes et al. (2000), Stock and Watson (2004), Timmermann (2006) and references therein. Forecast averaging has become so popular, with so many different ways to combine forecasts, that Andrawis et al. (2011) suggest to use hierarchical forecast combinations, i.e. combining combined forecasts.

While there is a large body of literature on forecasting day-ahead electricity prices and loads, only few of these studies examine the performance of combining forecasts obtained from individual models. To our best knowledge, existing applications so far only include the work by Bordignon et al. (2013), Løland et al. (2012), Raviv et al. (2013), Smith (1989), Taylor (2010) and Taylor and Majithia (2000). Hereby, earlier studies concentrate on forecasting electricity demand or transmission congestion, while only the two most recent focus on forecasting electricity spot prices.

Smith (1989) combines several ARIMA time series models for electricity demand. Hereby, the selection and combination of forecasts from different prediction methods is conducted on the basis of recent forecasting performance only, with no a priori assumptions about demand behavior. The author argues that such a procedure allows the prediction process to automatically adapt to the changing nature of the demand series over different days of the week and over different seasons. The weights of the forecast combinations change for different days of the week, to overcome cyclic modeling weaknesses of the individual models. Results of the empirical analysis suggest that the combined forecasts are significantly more accurate than any of the forecasts obtained from the individual models.

Taylor and Majithia (2000) apply switching and smooth transition forecast combination models for electricity demand profiling. The applied models allow for combining weights to vary across half-hourly intervals which is an appealing feature as different forecast models may be more suitable for different periods. A number of criteria are used to control the changing weights, including weather and the shape of demand profiles. Empirical results suggest an improved post-sample forecasting performance of the applied models.

Taylor (2010) applies so-called triple seasonal methods for short-term electricity demand forecasting. Hereby, double seasonal ARMA and two different double exponential smoothing methods are extended to accommodate an additional intra-year seasonal cycle. The author

illustrates the superior performance of the developed models over double seasonal methods. A result in particular relevant for our study is that further improvement in accuracy of the day-ahead forecasts is produced by using a combination of the forecasts from two of the applied methods.

Løland et al. (2012) forecast hourly day-ahead transmission congestion in the southern Norway (NO1) price area of the Nord Pool system. They utilize a number of prediction methods, including naive, exponential smoothing, ARIMA and TAR models. These forecasts are further combined, using equal weights (see Section 4.2.1), a weighted average with respect to estimated prediction error variances or a variant of the Bates and Granger (1969) averaging. The authors report that the latter approach yields the best results overall and that equal weights averaging is the worst of the three. However, for high absolute values of the net capacity utilization (i.e. a measure of transmission congestion), combining does not beat the naive approach.

Raviv et al. (2013) model each hourly price by considering the intra-day relation between the individual hours in the Nord Pool spot market. For the univariate analysis, they use heterogeneous autoregressive (HAR) and dynamic ARX models. For the multivariate analysis, they use VAR-type, Bayesian VAR, reduced rank regression (RRR), principal component regression and reduced rank Bayesian VAR models. The authors do not focus on investigating the usefulness of averaging forecasts, but in an empirical application they find that additional gain is achieved by using forecast combinations of individual models: even the best individual model is outperformed by forecast averaging (though not by a huge margin).

Finally, a recent paper by Bordignon et al. (2013) is the one most related to ours. Actually, our paper can be considered as an extension of their study. The authors combine five models: ARMAX, linear regression, time-varying regression and two variants of Markov regime-switching (MRS) to forecast British day-ahead electricity prices from five representative half-hourly load periods. All individual models are estimated recursively on an expanding window (as in this study), except for the MRS model which is additionally estimated using a rolling window of 6 months. Bordignon et al. (2013) consider five forecast averaging methods: equal weights (see Section 4.2.1), Bates-Granger (similar to IRMSE averaging, see Section 4.2.4), adaptive Bates-Granger and two variants of the adaptive Aggregated Forecast Through Exponential Re-weighting (AFTER) combination. The first two are calibrated using fixed 20- or 64-day selection periods, while the adaptive methods use either rolling or expanding windows, like in our study.

In their analysis, Bordignon et al. (2013) examine whether forecast combinations outperform individual methods, both from an ex post, i.e. using full sample information, and a much more realistic ex ante perspective, i.e. using only information available at the time the forecast is made. In the ex post analysis, they find that only a few cases (9%) are significant, although most combined forecasts (76%) perform better than individual forecasts. From the ex ante perspective, they find that again combined forecasts perform better (79%) than individual forecasts but this time as many as 33% cases are significant. On the other hand, only in 1% of the cases individual forecasts are significantly less accurate than the combined forecasts. While our findings also support the additional benefit of combining forecasts for deriving more accurate predictions, they are not as clear-cut. Using 12 individual models, four datasets from three different markets, longer calibration and out-of-sample time periods and, most importantly, seven different averaging schemes, we find that the performance is not uniform across the considered markets.

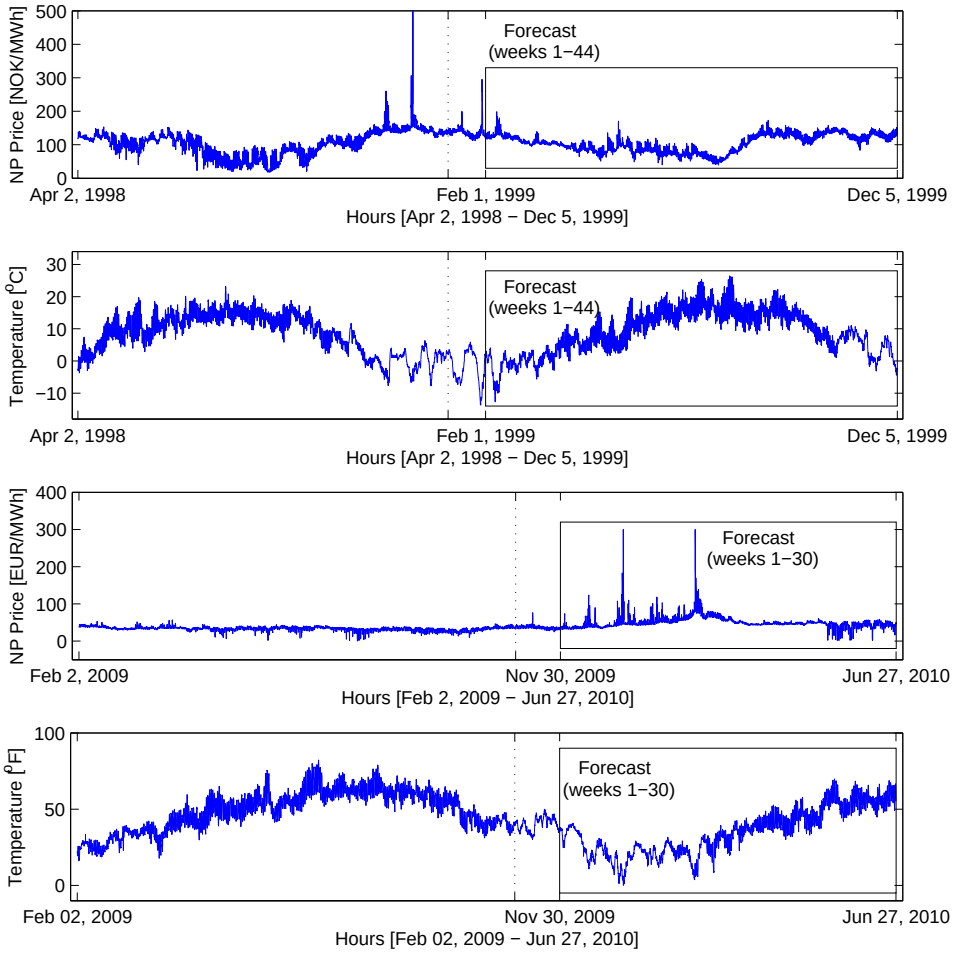

Figure 1: *Two upper panels*: Nord Pool hourly system prices (in NOK/MWh) and hourly air temperatures (in Celsius) for the period April 2, 1998 – December 5, 1999. *Two lower panels*: Nord Pool hourly system prices (in EUR/MWh) and hourly air temperatures (in Fahrenheit) for the period February 2, 2009 – June 27, 2010. In all subplots the out-of-sample test ('Forecast') periods are indicated by rectangles, while the vertical dotted lines represent the beginning of the calibration windows for the forecast averaging methods (four weeks prior to the test periods).

3. The datasets

The datasets used in this empirical study include four spot (or day-ahead) price time series from three major power markets: Nord Pool (NP; years 1998-1999 and 2009-2010), the European Energy Exchange (EEX; 2009-2011) and the Pennsylvania-New Jersey-Maryland Interconnection (PJM; 2010-2012). The oldest dataset (NP 1998-1999) was also studied by Weron and Misiorek (2008); the forecast averaging results (see Table 1) can be easily compared with those of the individual models (Table 2 in the cited paper). The three recent datasets provide a more timely description of price behavior. In contrast to previous studies on combining electricity spot price forecasts (Bordignon et al., 2013; Raviv et al., 2013), the range of data we use here allows for a thorough evaluation of the models under different conditions stemming from the different time periods, geographical areas and the different exchanges (frequency and severity of spikes, weather conditions, generation stack, market regulations, etc.).

3.1. NP99 – Nord Pool (1998-1999)

This dataset comprises hourly Nord Pool market clearing prices (in NOK/MWh) and hourly temperatures (in Celsius) for the period April 2, 1998 – December 5, 1999. The time series were constructed using data published by the Nordic power exchange Nord Pool (www.nordpool.com) and the Swedish Meteorological and Hydrological Institute (www.smhi.se) and preprocessed to account for missing values and changes to/from the daylight saving time (like in Weron, 2006, Section 4.3.7). The missing data values and a few outliers (e.g. an extremely low price surrounded by 4-5 times higher prices or a twice lower temperature figure than normal) were substituted by the arithmetic average of the two neighboring values. The 'doubled' values (corresponding to the changes from the daylight saving/summer time) were substituted by the arithmetic average of the two values for the 'doubled' hour.

The air temperature was chosen as the exogenous (fundamental) variable; typically it is the most influential on electricity prices weather variable (Weron, 2006). The actual temperatures observed on day $T+1$ were used as the 24 hourly day-ahead temperature forecasts available on day T. Slightly different – possibly better – results would be obtained if day-ahead temperature forecasts were used, but these were not available to us.

Following earlier studies, see e.g. Conejo et al. (2005) and Misiorek et al. (2006), we applied the logarithmic transformation to the price series in order to attain a more stable variance and removed the mean log-price and the median temperature to center the data around zero. The dependence between log-prices and temperatures is moderately anticorrelated, i.e. low temperatures in Scandinavia imply high electricity prices at Nord Pool and vice versa, see the top two panels in Figure 1; Weron and Misiorek (2008) report that in the studied period the Pearson correlation between log-prices and temperatures is negative ($\rho = -0.47$) and significant (p-value ≈ 0; null of no correlation). The 'hourly air temperature' is in fact a proxy for the air temperature in the whole Nord Pool region. It is calculated as an arithmetic average of the hourly air temperatures in six Scandinavian cities/locations (Bergen, Helsinki, Malmö, Stockholm, Oslo and Trondheim).

Finally, note that in comparison to the study of Weron and Misiorek (2008), the forecast period has been significantly extended. Instead of considering four five-week test periods corresponding to the seasons of the year, we analyze a 44 weeks long forecast window (February 1, 1999 – December 5, 1999). Weeks 1-5 correspond to period II (February) in Weron and Misiorek (2008), weeks 13-17 to period V (May), weeks 27-31 to period VIII (August) and weeks 40-44 to period XI (November). Like in the cited paper, the calibration window starts on April 2, 1998.

3.2. NP10 – Nord Pool (2009-2010)

This dataset was constructed similarly to the previous one. It comprises Nord Pool hourly system prices (in EUR/MWh) and hourly air temperatures (in Fahrenheit) for the period February 2, 2009 – June 27, 2010. However, this time the 'hourly air temperature' was calculated as an arithmetic average of the observed hourly air temperatures in five Scandinavian cities/locations: Copenhagen, Helsinki, Oslo, Stockholm and Trondheim using data provided by NOAA's National Climatic Data Center (www.ncdc.noaa.gov).

The logarithms of hourly temperatures T_t (now in Fahrenheit, not Celsius) were used as the exogenous variable in the time series models for the log-prices. More precisely, we applied the following transform: $\log(T_t - \min T_t + 1)$, so that the obtained series is bound by zero from below. A limited forecasting study we performed has shown that temperatures transformed in such a way lead to more accurate spot price forecasts than the temperatures themselves. As

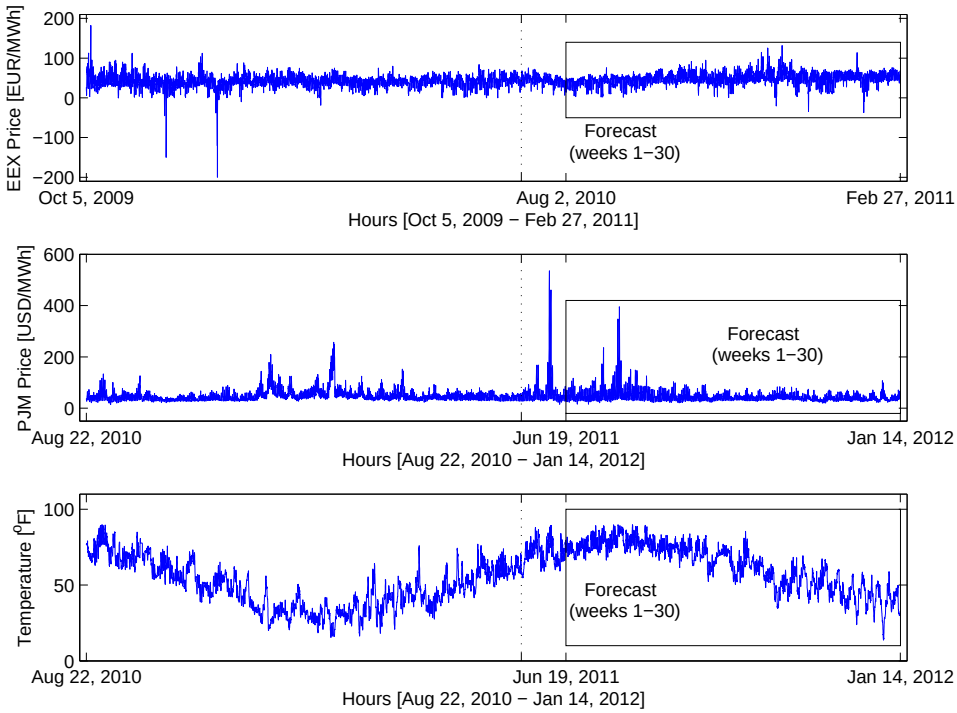

Figure 2: *Top panel*: EEX hourly system prices (in EUR/MWh) for the period October 5, 2009 – February 27, 2011; note that the EEX dataset is the only one without a fundamental variable. *Middle and bottom panels*: PJM hourly system prices (in USD/MWh) and hourly air temperatures (in Fahrenheit) for the period August 22, 2010 – January 14, 2012. In all subplots the out-of-sample test ('Forecast') periods are indicated by rectangles, while the vertical dotted lines represent the beginning of the calibration windows for the forecast averaging methods (four weeks prior to the test periods).

before, the mean log-price and the median log-temperature were removed to center the data around zero. The forecast period consists of 30 weeks, see the two lower panels in Figure 1. It covers the period November 30, 2009 – June 27, 2010. The calibration window starts on February 2, 2009.

3.3. EEX – European Energy Exchange (2009-2011)

The last European dataset comprises EEX hourly system prices (in EUR/MWh) for the period October 5, 2009 – February 27, 2011 (source: www.eex.com). They were preprocessed for missing, 'doubled' values and outliers in a way similar to that for the previous datasets. Because negative electricity prices were recorded during the studied period (see the top panel in Figure 2), no logarithm transform was used. This dataset is also the only one without a fundamental (exogenous) variable. As a consequence, the number of considered time series models is reduced by a half. Like for the NP10 and PJM datasets, the forecast period consists of 30 weeks. It covers the period August 2, 2010 – February 27, 2011. The calibration window starts on October 5, 2009.

3.4. PJM – Pennsylvania-New Jersey-Maryland Interconnection (2010-2012)

The last dataset comprises hourly day-ahead locational marginal prices (LPMs) for the PJM JCPL zone and hourly air temperatures for New York City. The time series were constructed using data published by EDF Suez (www.gdfsuezenergyresources.com) and NOAA's National

Climatic Data Center (www.ncdc.noaa.gov). They were preprocessed for missing, 'doubled' values and outliers in a way similar to the other datasets.

The logarithms of hourly temperatures T_t (in Fahrenheit) were used as the exogenous variable in the time series models for the log-prices (no additional transformations were required since the temperatures were above 1^oF in the studied period). This selection was motivated by a roughly linear dependence between these two variables. The mean log-price and the median log-temperature were removed to center the data around zero. The forecast period consists of 30 weeks and covers the period June 19, 2011 – January 14, 2012, see the two lower panels in Figure 2. The calibration window starts on August 22, 2010.

4. Individual models and averaging schemes

In all individual models the weekly seasonal behavior is captured by a combination of the autoregressive structure of the models and daily dummy variables. Note that unlike in derivatives pricing and risk management applications where daily average spot prices are typically used (see e.g. Bierbrauer et al., 2007; Janczura et al., 2013), in day-ahead forecasting of hourly electricity prices the long-term trend-seasonal component is usually not taken into account as it adds unnecessary complexity to the already parameter-rich models (Conejo et al., 2005; Nogales et al., 2002; Weron and Misiorek, 2008). However, if the price series is decomposed into the trend-seasonal and stochastic components, then the former has to be assumed to be known ex ante (like in Bordignon et al., 2013) or predicted (for a recent review, see Nowotarski et al., 2013), and added back before computing the forecasting errors.

All models and all averaging schemes are estimated using an expanding window incorporating all possible information up to the point the forecast is made. For instance, to forecast Nord Pool prices for the 24 hours of February 1, 1999 we use prices and temperatures from the period April 2, 1998 – March 31, 1999. Next, to forecast the prices for the 24 hours of February 2 we use prices and temperatures from the period April 2, 1998 – February 1, 1999 and so forth. For the averaging schemes we also require a calibration period in order to determine the optimal model weights. We decided to use a period of four weeks, where forecasts from individual models are already available, for the initial calibration of the averaging methods, see the dotted vertical lines in Figures 1 and 2.

4.1. Individual models

A typical and obvious caveat shared by all empirical applications using forecast averaging is that results depend on the specific choice of individual models. Eliminating this effect by using all conceivable models is unreasonable, if not impossible. Thus, our choice of individual models is guided by previous literature. In particular, we utilize a set of six carefully selected model classes that have been analyzed by Weron and Misiorek (2008): AR/ARX models, spike preprocessed AR/ARX models (where the model structure was estimated after replacing price spikes with less extreme observations), threshold AR/ARX models, mean-reverting jump diffusion models, and two classes of semiparametric AR/ARX models.

The modeling was implemented separately across the hours, leading to 24 sets of parameters for each day the forecasting exercise was performed. This approach was inspired by the fact that each hour displays a rather distinct price profile, reflecting the daily variation of demand, costs and operational constraints, and by the extensive research on demand forecasting, which has generally favored the multi-model specification for short-term predictions (see e.g. Bunn, 2000; Shahidehpour et al., 2002; Weron, 2006).

In the following paragraphs we briefly review the time series models we use for generating individual model forecasts. We refer the reader to Weron and Misiorek (2008) to find a complete description of these models and a thorough discussion of the model choices.

4.1.1. AR and ARX models

The basic autoregressive model structure used in the study is given by the following formula (denoted later in the text as **ARX**):

$$\begin{aligned} p_t &= \phi_1 p_{t-24} + \phi_2 p_{t-48} + \phi_3 p_{t-168} + \phi_4 mp_t + \\ &\quad + \psi_1 z_t + d_1 D_{Mon} + d_2 D_{Sat} + d_3 D_{Sun} + \varepsilon_t. \end{aligned} \tag{1}$$

The lagged log-prices p_{t-24}, p_{t-48} and p_{t-168} account for the autoregressive effects of the previous days (the same hour yesterday, two days ago and one week ago), while mp_t creates the link between bidding and price signals from the entire previous day (it is the minimum of the previous day's 24 hourly log-prices). The variable z_t refers to the actual hourly temperature (Nord Pool 1998-1999) or the logarithm of the hourly temperature (Nord Pool 2009-2010, PJM 2011-2012). Recall that for the EEX dataset no fundamental variable is used. The three dummy variables – D_{Mon}, D_{Sat} and D_{Sun} (for Monday, Saturday and Sunday, respectively) – account for the weekly seasonality. Finally, the ε_t's are assumed to be independent and identically distributed (i.i.d.) with zero mean and finite variance. Restricting the parameter $\psi_1 = 0$ yields the **AR** model. Model parameters were estimated in Matlab by minimizing the Final Prediction Error (FPE) criterion.

4.1.2. p-AR and p-ARX models

Linear models like AR and ARX are sensitive to outliers, i.e., extreme observations that deviate significantly from the 'usual' values. Although we do not believe that forecasters should ignore the electricity price spikes, for the purpose of predicting the mean level of next day's prices we could follow a relatively popular approach of substituting the spikes with 'less unusual' values (Conejo et al., 2005; Nogales et al., 2002; Shahidehpour et al., 2002; Weron, 2006). This can be done in a number of ways. Here we use the 'damping scheme', where an upper limit T^* is set on the price (equal to the mean plus three standard deviations of the price in the calibration period) and all prices exceeding T^* are set to $P_t = T^* + T^* \log_{10}(P_t/T^*)$. The spike preprocessed models, denoted in the text as **p-ARX** and **p-AR**, also utilize formula (1), with the only difference that the data used for calibration is spike preprocessed using the damping scheme.

4.1.3. TAR and TARX models

As an alternative to 'damping' or eliminating price spikes we can use a time series model that allows for changes of regime (or price behavior), like the Threshold Auto Regressive (TAR) model of Tong and Lim (1980). In such a model the regime switching between two (or more, in general) autoregressive processes is governed by the value of an observable threshold variable v_t relative to a chosen threshold level T_0. So the two additional models: **TAR** and **TARX** are a two-state extension of (1) with $T_0 = 0$ and v_t equal to the difference in mean prices for yesterday and eight days ago. Like for the autoregressive models, the parameters can be estimated by minimizing the FPE criterion.

4.1.4. MRJD and MRJDX models

The next pair of models is based on a discrete-time version of a mean-reverting jump diffusion process. The **MRJDX** (**MRJD** when $\psi_1 = 0$) specification used in this study is given by the following formula:

$$p_t = \phi_1 p_{t-24} + \psi_1 z_t + d_1 D_{Mon} + d_2 D_{Sat} + d_3 D_{Sun} + \varepsilon_{t,i}, \tag{2}$$

where the subscript i takes the value of 1 when there is no jump or 2 if there is a jump, $\varepsilon_{t,1} \sim N(0, \sigma^2)$ and $\varepsilon_{t,2} \sim N(\mu, \sigma^2 + \gamma^2)$. The model can be easily estimated by maximum likelihood, with the likelihood function being a product of the densities of a mixture of two normals (Ball and Torous, 1983).

4.1.5. IHMAR, IHMARX, SNAR and SNARX models

Finally, we use two additional more flexible AR specifications: the iterated Hsieh-Manski estimator (**IHMAR**) and the smoothed nonparametric ML estimator (**SNAR**). These models relax the normality assumption needed for the ML estimation in the standard AR model. We keep the functional AR form but obtain the estimates from a numerical maximization of the empirical likelihood as applied in Hsieh and Manski (1987), Cao et al. (2003) and in our context of electricity price forecasting, in Weron and Misiorek (2008) where a more detailed description is found. The corresponding models with the additional exogenous variables are denoted as **IHMARX** and **SNARX**.

4.2. Forecast averaging techniques

As mentioned above, we are interested in how well the approach of combining forecasts provided by individual models performs in the context of spot electricity prices. Given the promising results of Bordignon et al. (2013) and Raviv et al. (2013), we aim to use a variety of forecast averaging techniques in order to thoroughly examine the performance of these methods. In the following, we provide a description of the averaging approaches being applied in the empirical analysis.

The M individual log-price forecasts $\widehat{p}_{1t}, \ldots, \widehat{p}_{Mt}$ calculated for time t are inverted back to levels $\widehat{P}_{1t}, \ldots, \widehat{P}_{Mt}$ and the combined spot price forecast is given by:

$$\widehat{P}_t^c = \sum_{i=1}^{M} w_{it} \widehat{P}_{it}, \tag{3}$$

where w_{it} is the weight assigned at time t to forecast i. We calculate the weights recursively at each time step, using data covering the first prediction point (indicated by the dotted vertical lines in Figures 1 and 2) until $t - 24$ (i.e. 24 hours prior to the hour we forecast the price for).

4.2.1. Simple averaging

The most natural approach to forecast averaging is the use of the (arithmetic) mean of all forecasts produced by the different models. It is highly robust and is widely used in business and economic forecasting, see e.g. Clemen (1989), Stock and Watson (2004) and Genre et al. (2013). In the following, we denote this approach as **Simple**.

4.2.2. OLS and LAD averaging

Ordinary Least Squares (**OLS**) averaging is another easy-to-implement approach with a good performance track-record. The idea was first described in Crane and Crotty (1967), but it was the influential paper of Granger and Ramanathan (1984) to inspire more research effort in this direction. Since then, OLS averaging took on numerous variations according to different findings from different datasets.

In the the original proposal the combined forecast is determined using the following regression:

$$P_t = w_{0t} + \sum_{i=1}^{M} w_{it}\widehat{P}_{it} + e_t. \tag{4}$$

Thus, the corresponding electricity price forecast combination $\widehat{P}_t^c$ at time t using M models, is calculated as

$$\widehat{P}_t^c = \widehat{w}_{0t} + \sum_{i=1}^{M} \widehat{w}_{it}\widehat{P}_{it}. \tag{5}$$

This has the advantage of generating unbiased combined forecasts without the need to question individual models bias, and, of course, the use of OLS carries all its good properties. However, a few issues remain. It is natural to assume that different forecasts for the same target will be correlated. This means that the vector of estimated weights $\mathbf{w}_t$ is likely to exhibit an unstable behavior, a problem sometimes dubbed 'bouncing betas'. As a result, minor fluctuations in the sample can cause major shifts of the weight vector.

To address this issue we propose here a more robust to outliers version of (4) for forecast averaging. That is, we apply the absolute loss function $\sum_t |e_t|$ instead of the quadratic loss function $\sum_t {e_t}^2$ in (4) to yield the Least Absolute Deviation (**LAD**) regression. A possible advantage of using this absolute loss function is that it is more robust to electricity price spikes. Consider, for example, a model that performs well in general, yet sharply misses on specific dates, for example, during a period with a price spike. Using the quadratic loss function leads to a relatively large decrease of this model's weight, while using the absolute loss function may yield a relatively smaller decrease of the weight.

The LAD regression may be viewed as a special case of quantile regression which allows to develop explicit models for specific quantiles of the distribution of the dependent variable (Koenker, 2005). In energy economics, quantile regression has been applied to forecasting Value-at-Risk levels (Bunn et al., 2013) and computing predictive densities for day-ahead electricity prices (Jonsson et al., 2013), but not in the context of forecast averaging. Taking the quantile to be 0.5 (i.e. the median) simply yields the LAD regression.

4.2.3. PW and CLS constrained averaging

Another issue in OLS averaging is the interpretation of the results. It is hard to explain a linear combination with negative weights for some models, which is likely to result from this approach. Therefore, we follow Sevket and Aksu (1989) and apply two more variants of (4) using additional constraints on the estimated weights. First, we constrain the estimated coefficients to some pre-defined regions, allowing only for positive weights (**PW**):

$$w_{0t} = 0 \quad \text{and} \quad w_{it} \geq 0, \qquad \forall i, t. \tag{6}$$

Restricting weights to be non-negative was found by Aksu and Sevket (1992) to be a strong competitor to the robust simple average and to almost always outperform the unconstrained OLS.

Secondly, we use an additional condition forcing the weights to sum up to one using constrained least squares (**CLS**) estimation:

$$w_{0t} = 0 \quad \text{and} \quad \sum_{i=1}^{M} w_{it} = 1, \qquad \forall t. \tag{7}$$

Note that the first option allows individual forecasts to be biased, since the sum $\sum_{i=1}^{M} w_{it}$ does not necessarily sum to one. That said, when the forecasts are unbiased then $\sum_{i=1}^{M} w_{it}$ should not deviate substantially from one which holds in this case as well. Using (7), we gain a natural interpretation of the coefficients w_{it} which can be viewed as relative importance of each model in comparison to all other models. Note that there is no closed form solution for the constrained models PW and CLS, however, they can be solved using quadratic programming.

4.2.4. IRMSE averaging

Another performance-based approach is to choose the weights for each model based on the inverse of the Root Mean Squared Errors (RMSE). Clearly, using this approach, models producing smaller RMSE will be assigned larger weights in comparison to models with higher RMSE. A similar approach has been suggested, for example, by Diebold and Pauly (1987) and has been successfully applied by Stock and Watson (2004) for combining forecasts of output growth. Overall, using the approach, model weights are determined based on their forecasting performance and can be defined as:

$$w_{it} = \frac{\left(\frac{RMSE_{it}}{\sum_{i=1}^{M} RMSE_{it}}\right)^{-1}}{\sum_{i=1}^{M}\left(\frac{RMSE_{it}}{\sum_{i=1}^{M} RMSE_{it}}\right)^{-1}} = \frac{\frac{1}{RMSE_{it}}}{\sum_{i=1}^{M}\frac{1}{RMSE_{it}}}. \tag{8}$$

Here, $RMSE_{it}$ denotes the out-of-sample performance for model i and is computed in a recursive manner using forecast errors from the first prediction point up to $t - 24$ hours. We denote this method in the text as **IRMSE**.

4.2.5. Bayesian Model Averaging (BMA)

So far we have considered forecast combination schemes which apply weights to all considered models. The idea of Bayesian Model Averaging (denoted as **BMA**) is to relax this assumption and avoid the a priori decision to use all models. Such an approach will result in a substantial increase of possible forecast combinations, 2^M to be exact. Even with a moderate number of models, for example $M = 50$ such an approach is infeasible. However, in this study we have $M = 12$ models (apart from the EEX dataset with $M = 6$), making the process very accurate albeit slow. Accurate since we can compute *all* possible models (this is done using an efficient branch-and-bound algorithm implemented in R's BMA package, see Raftery et al., 2005). The model weights for BMA are given by the Bayes' theorem where we compute the posterior probabilities for each of the individual options:

$$w_{lt} = \frac{L(P_t|m_l, D)\rho(m_l)}{\sum_{j=1}^{2^M} L(P_t|m_j, D)\rho(m_j)}. \tag{9}$$

Note that m does not stand here for a particular individual model (e.g. AR), but for a model combination option, i.e. one of the 2^M available options. Then $L(p_t|m_l, D)$ denotes the likelihood of the price P_t given option m_l and data D. We do not use the subscript i but l to reflect

the fact that $\boldsymbol{w}$ is not the weight vector for each model but a very long (i.e. of length 2^M) weight vector for each combination of models. Note that similar to the other approaches, D is the data up to time $t-24$. In order to estimate equation (9), a prior distribution over each option $\rho(m_l)$ is required. Following the most common practice, we use a uniform prior over all options, acknowledging our uncertainty with regard to the true combination of models. Madigan and Raftery (1994) suggest that averaging over all the models in this fashion on average provides a better predictive ability than any single model in the pool. Once the weights are set, the conditional expectation of the forecast is calculated for each of the considered options, and the resulting forecast combination is given by:

$$\widehat{P}_t^c = \sum_{l=1}^{2^M} w_{lt} \mathbb{E}(P_t | m_l, \theta_l). \tag{10}$$

Here $\boldsymbol{\theta}_l$ is the collection of parameters required for combination option l. For example, if the combination option l is the AR model and the TAR model with zeros for the rest, then θ_l are the estimated parameters for these two models.

4.2.6. Best Individual (BI) model selection and the benchmark model

Next to the suggested forecast combination models, it is of interest to examine the performance of the best individual model. In a straightforward manner the best individual model could be defined as the best performing individual model from an ex post perspective (as in Bordignon et al., 2013). Although theoretically pleasing, an ex post analysis is not feasible in practice – one cannot use information observed only at time $T+1$ for forecasting conducted at time T. Hence in this paper we investigate how well the approach of combining forecasts performs versus the realistic alternative of selecting a single model specification beforehand. We select the ARX model (or the AR model for the EEX dataset) as a relatively simple, yet robust, benchmark. This particular ARX model specification, see formula (1), has been shown to perform very well for the California market (Misiorek et al., 2006; Weron, 2006). However, ARX models in general, sometimes referred to as 'dynamic regressions', were also found to yield very good day-ahead price predictions for Nord Pool (Raviv et al., 2013), PJM (Conejo et al., 2005), and the UK (Gonzalez et al., 2012) and Spanish markets (Nogales et al., 2002).

In another way, we can also consider the best individual ex ante model (**BI**), i.e. making the decision to pick one of the models that performed best in the past, and examine its future performance from time t onwards. The BI is essentially a model (or forecast) selection scheme, but we can view it also as a special case of forecast averaging with degenerate weights given by the vector:

$$w_{it} = \begin{cases} 1 & \text{if model } i \text{ achieves lowest forecast error during the calibration period,} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We decide to choose the BI model as the one yielding the best forecasts in terms of RMSE for the data covering the first prediction point up to $t-24$ hours, like for the forecast averaging schemes. Note that in what follows we use a *weekly* evaluation metric as in Weron and Misiorek (2008), while the weights for the individual forecasts and in particular the choice of BI are determined on a daily basis.

5. Empirical results

We now present out-of-sample forecasting results for the considered datasets. We examine day-ahead forecasts of hourly market clearing prices for the following markets and periods: (i) the Nord Pool market for 44 weeks from February 1, 1999 to December 5, 1999 as well as (ii) during the 30 week period November 30, 2009 – June 27, 2010; (iii) the EEX market during the 30 week period from August 2, 2010 to February 27, 2011; and (iv) the PJM market for a 30 week period from June 19, 2011 to January 14, 2012. For further information on applied calibration windows for each of these markets see Section 3.

Forecasts for the considered models are determined the following way: models (as well as model parameters and combination weights) are reestimated on a daily basis and a forecast for all 24 hours of the next day is determined at the same point in time. Forecasts are first calculated for each of the 12 individual models (or six for the EEX dataset) and then combined according to estimated weights for each of the seven forecast averaging approaches and one model selection scheme.

Following Conejo et al. (2005) and Weron and Misiorek (2008), we compare the methods in terms of the Weekly-weighted Mean Absolute Error (WMAE) loss function and evaluate the forecast performance using weekly time intervals, each with $24 \times 7 = 168$ hourly observations. Note that we also analyzed the forecasts using squared error losses, however, results were qualitatively similar and are omitted here due to space limitations. For each week we calculate the WMAE for averaging method i as:

$$\mathrm{WMAE}_i = \frac{1}{\overline{P}_{168}}\mathrm{MAE}_i = \frac{1}{168 \cdot \overline{P}_{168}} \sum_{h=1}^{168} \left|P_h - \hat{P}_{ih}\right|, \tag{12}$$

where P_h is the actual price for hour h (not the log-price p_h), $\hat{P}_{ih}$ is the predicted price for that hour obtained for averaging method i and $\overline{P}_{168} = \frac{1}{168}\sum_{h=1}^{168} P_h$ is the mean price for a given week. The WMAE is similar to the Mean Absolute Percentage Error (MAPE) with P_h replaced by $\overline{P}_{168}$ in the denominator of (12). This is done to avoid the adverse effect of prices close to zero.

Overall, we provide results for eight different approaches as well as for the benchmark ARX (or AR for the EEX market) model:

1) Simple – a simple average of the forecasts provided by all 12 (or six for EEX) individual models,

2) OLS – forecast combination with weights determined by eqn. (4) using standard OLS,

3) LAD – forecast combination with weights determined by eqn. (4) using least-absolute-deviation regression,

4) PW – forecast combination with weights determined by eqn. (4) only allowing for positive weights $w_{it} \geq 0$,

5) CLS – forecast combination with weights determined by eqn. (4) with constraints $w_{it} \geq 0$ and $\sum_{i=1}^{M} w_{it} = 1$,

6) IRMSE – forecast combination with weights determined by eqn. (8),

7) BMA – forecast combination with weights determined using Bayesian Model Averaging,

8) BI – the individual method that would have been chosen ex ante, based on its forecasting performance from the first prediction point until $t - 24$ hours earlier.

5.1. Performance evaluation

In Tables 1–4 we report the WMAE for each week in the forecasting period (rectangles in Figures 1 and 2) and for the seven forecast averaging (FA) schemes, one model selection technique (BI) and the benchmark model (ARX or AR). Note that we use the term *model* to describe both individual models and averaging schemes. Summary statistics are presented in the bottom rows and include: $\overline{\text{WMAE}}$ – the mean value of WMAE for a given model (with standard deviation in parentheses), *# better* – the number of weeks a given averaging method is better than the benchmark or the BI (in separate rows), *# best* – the number of weeks a given averaging method performs best in terms of WMAE (only among the eight forecast averaging and selection schemes, i.e. excluding the benchmark and other individual models) and *m.d.f.b.* – the mean deviation from the best model (averaging or individual) in each week. The latter measure indicates how similar is a models' performance to the 'optimal model' composed of the overall best performing (individual or combined) model in each week. Note that we do not report here the WMAE errors of all individual models, only of the benchmark ARX (or AR) model. The full error tables may be obtained from the authors upon request.

Let us first examine the results for the first forecasting period from the Nord Pool market. It comprises 44 weeks from February 1, 1999 to December 5, 1999, see Table 1. During this period all averaging techniques outperform the benchmark ARX model with respect to WMAE and m.d.f.b. The best results are obtained for LAD and PW where the WMAE is reduced by approximately 10% in comparison to the benchmark model, while the mean deviation from the best model is reduced by even more than 40% over the considered 44 week period. All forecast averaging techniques provide better results than the ARX model for at least 29 weeks, while the PW approach outperforms the benchmark model for 34 out of the 40 weeks. Among the eight forecast and selection schemes, LAD yields the best results for more than 25% of the time, providing the smallest WMAE for 13 of the 44 weeks.

With respect to the other benchmark, the BI selection approach, we get the following results: while BI performs significantly better than the benchmark ARX model, we find that all but one FA techniques outperform BI with respect to the considered WMAE and m.d.f.b. criteria. Also, all forecast averaging techniques provide better results than BI for at least 25 out of 40 weeks. Only the OLS approach fails to consistently outperform BI, but still provides results of the same quality. Overall, our results indicate that for the considered period, combining forecasts clearly seems to outperform the benchmark ARX model, but also the BI approach, with respect to the considered loss function and measures.

Table 2 provides out-of-sample forecasting results from Nord Pool market for the more recent period which covers 30 weeks from November 30, 2009 to June 27, 2010. Also for this dataset, the majority of averaging techniques clearly outperforms the ARX model with respect to the WMAE and m.d.f.b. criteria. Only OLS and BMA averaging yield results that are slightly worse than for the benchmark approach. For all other FA techniques the average WMAE is reduced by 9%-17%, while the mean deviation from the best model is reduced by 37%-72% over the considered 30 week period. Clearly, the best results are obtained for the forecast combinations generated by the least-absolute-deviation (LAD) regression. For 26 out of the 30 weeks considered, LAD performs better than the ARX model, while it yields the overall best results of all models for 13 weeks. LAD also provides the smallest average WMAE and m.d.f.b., followed by simple averaging and IRMSE.

Table 1: *NP 1998-1999 dataset (NP99)*. WMAE for the seven forecast averaging schemes, one model selection technique (BI) and the benchmark model (ARX). Summary statistics are provided in the bottom rows: $\overline{\text{WMAE}}$ is the mean value of WMAE for a given model (with standard deviation in parentheses), # *better* is the number of weeks a given averaging method is better than the benchmark or the BI, # *best* is the number of weeks a given averaging method performs best in terms of WMAE (i.e. excluding the individual methods), and finally *m.d.f.b.* is the mean deviation from the best model (averaging or individual) in each week. Emphasized in bold are the WMAE, $\overline{\text{WMAE}}$ and m.d.f.b. values lower than those of the benchmark. The out-of-sample period ranges from February 1 to December 5, 1999.

Week	Simple	OLS	LAD	PW	CLS	IRMSE	BMA	BI	ARX
1	**4.15**	**3.25**	**2.83**	**4.49**	**4.59**	**4.22**	**3.17**	**4.17**	4.63
2	**3.07**	3.89	**3.27**	**3.01**	**3.28**	**3.07**	3.86	**3.31**	3.59
3	**2.82**	**2.33**	**2.01**	**2.62**	**3.12**	**2.90**	**2.25**	**3.24**	3.65
4	**3.58**	**3.43**	**3.22**	**2.81**	**3.99**	**3.69**	**3.51**	**4.41**	4.85
5	**4.22**	**3.44**	**2.90**	**3.10**	**4.17**	**4.34**	**3.67**	**4.55**	5.63
6	**3.36**	**3.65**	**3.28**	**3.33**	**3.30**	**3.40**	**3.71**	**3.37**	3.94
7	**3.22**	**2.47**	**1.76**	**2.38**	**2.93**	**3.31**	**2.51**	**3.16**	4.29
8	**2.58**	**1.51**	**1.21**	**1.52**	**2.27**	**2.63**	**1.50**	**2.56**	3.54
9	**4.76**	**2.88**	**2.63**	**3.31**	**4.01**	**4.83**	**2.89**	**4.31**	5.68
10	**7.16**	**6.16**	**6.02**	**6.18**	**6.81**	**7.18**	**6.21**	**6.87**	7.51
11	5.89	6.00	6.17	5.92	6.39	5.88	5.96	6.08	5.73
12	5.19	**4.13**	**4.51**	**4.36**	5.36	5.17	**4.07**	5.04	5.02
13	**4.29**	**3.62**	**3.62**	**3.66**	**3.97**	**4.30**	**3.61**	**4.00**	4.59
14	6.40	6.44	6.48	6.42	6.95	6.35	6.41	6.99	5.84
15	8.82	9.28	9.86	9.38	9.22	8.74	9.23	9.23	8.04
16	6.67	**5.75**	6.04	**5.68**	6.60	6.59	**5.67**	6.21	5.97
17	**4.96**	5.33	5.15	**4.83**	**4.95**	**4.95**	5.36	**5.03**	5.11
18	4.16	4.82	4.89	4.24	4.17	4.09	4.74	3.81	3.59
19	6.81	**5.99**	**6.28**	**5.87**	6.54	6.79	**5.89**	6.60	6.49
20	4.98	5.01	5.19	4.99	5.00	4.91	4.89	4.69	4.66
21	5.54	**4.65**	**4.59**	**4.67**	**5.32**	5.49	**4.59**	**5.33**	5.44
22	7.59	7.51	7.76	7.27	7.60	7.50	7.44	7.19	6.89
23	**5.81**	6.28	6.18	6.45	**5.78**	**5.75**	6.19	**5.82**	5.98
24	5.33	**4.20**	**3.82**	**3.73**	**4.84**	5.31	**4.10**	**4.88**	4.91
25	7.04	8.21	**6.60**	**5.37**	**6.65**	6.96	8.24	**6.69**	6.88
26	**4.79**	**5.19**	**4.76**	**5.23**	**4.79**	**4.84**	**5.25**	**5.26**	5.70
27	**3.33**	5.94	5.46	5.08	**3.72**	**3.37**	5.80	**4.29**	4.64
28	**4.88**	7.53	6.88	6.42	**5.09**	**4.89**	7.61	**5.44**	5.89
29	**4.38**	8.23	7.18	6.34	**4.64**	**4.36**	8.30	**5.13**	5.82
30	**4.25**	6.17	**5.46**	**5.42**	**4.34**	**4.30**	6.14	**5.10**	5.81
31	**2.85**	4.11	**3.35**	**3.62**	**2.94**	**2.83**	4.07	**3.15**	3.66
32	**1.87**	**2.53**	**2.20**	**2.27**	**1.92**	**1.89**	**2.47**	**2.25**	2.77
33	**2.84**	**3.18**	**3.03**	**3.19**	**2.74**	**2.83**	**3.21**	**3.01**	3.59
34	2.91	**2.70**	2.91	**2.64**	2.96	2.92	**2.69**	3.13	2.74
35	**2.35**	**2.31**	**2.20**	**2.39**	**2.40**	**2.38**	**2.32**	**2.37**	2.61
36	**4.36**	**4.52**	**4.42**	**4.62**	**4.13**	**4.31**	**4.46**	**4.36**	4.68
37	**3.54**	**3.50**	**3.54**	**3.71**	**3.40**	**3.57**	**3.50**	**3.56**	4.07
38	**2.39**	**2.36**	**2.32**	**2.46**	**2.32**	**2.40**	**2.37**	**2.47**	2.71
39	2.40	**2.22**	2.34	**2.22**	2.47	2.41	**2.22**	2.52	2.30
40	**2.73**	**2.47**	**2.36**	**2.47**	**2.76**	**2.79**	**2.47**	**2.80**	2.94
41	**3.49**	**3.37**	**3.54**	**3.61**	**3.38**	**3.55**	**3.42**	**3.65**	3.91
42	**2.55**	**2.54**	**2.52**	**2.54**	**2.47**	**2.54**	**2.53**	**2.66**	2.77
43	**2.22**	**1.79**	**1.77**	**1.95**	**2.25**	**2.24**	**1.75**	**2.16**	2.33
44	**3.16**	**3.04**	**3.23**	**3.30**	**2.94**	**3.21**	**3.10**	**3.30**	3.47
				Summary statistics					
$\overline{\text{WMAE}}$	**4.31**	**4.41**	**4.22**	**4.21**	**4.31**	**4.32**	**4.39**	**4.41**	4.66
	(1.66)	(1.93)	(1.92)	(1.69)	(1.68)	(1.63)	(1.92)	(1.60)	(1.44)
# better than AR	30	29	30	34	33	30	29	33	–
# better than BI	29	27	30	28	35	25	27	–	–
# best	4	1	13	7	5	6	5	3	–
m.d.f.b.	**0.68**	**0.78**	**0.59**	**0.57**	**0.67**	**0.69**	**0.76**	**0.78**	1.01

Table 2: *NP 2009-2010 dataset (NP10)*. WMAE for the seven forecast averaging schemes, one model selection technique (BI) and the benchmark model (ARX). Summary statistics are provided in the bottom rows: $\overline{\text{WMAE}}$ is the mean value of WMAE for a given model (with standard deviation in parentheses), *# better* is the number of weeks a given averaging method is better than the benchmark or the BI, *# best* is the number of weeks a given averaging method performs best in terms of WMAE (i.e. excluding the individual methods), and finally *m.d.f.b.* is the mean deviation from the best model (averaging or individual) in each week. Emphasized in bold are the WMAE, $\overline{\text{WMAE}}$ and m.d.f.b. values lower than those of the benchmark. The out-of-sample period ranges from November 30, 2009 to June 27, 2010.

Week	Simple	OLS	LAD	PW	CLS	IRMSE	BMA	BI	ARX
1	**5.11**	5.73	**4.61**	5.37	**5.13**	**5.15**	5.75	5.86	5.36
2	**2.97**	**2.56**	**1.91**	**2.29**	**2.56**	**3.00**	**2.38**	**3.32**	3.39
3	**14.38**	18.75	15.03	15.04	**14.53**	**14.34**	18.90	15.60	14.98
4	**6.04**	9.66	**5.03**	7.81	**6.72**	**6.04**	9.09	**6.95**	7.66
5	**4.16**	7.45	**3.58**	5.63	5.10	**4.16**	7.35	**4.57**	4.62
6	**28.73**	**31.48**	**27.31**	**30.53**	**30.28**	**28.74**	**31.20**	31.59	31.50
7	10.58	12.24	10.25	10.16	10.41	10.60	12.33	10.24	9.92
8	5.18	5.27	**4.08**	**3.49**	**4.88**	5.16	5.44	4.97	4.97
9	14.95	**12.32**	**13.92**	**12.71**	**14.69**	14.97	**12.35**	**14.69**	14.91
10	**7.44**	**7.70**	**6.35**	**6.30**	**7.97**	**7.43**	**7.85**	**8.03**	8.60
11	**7.66**	**6.17**	**5.57**	**5.56**	**8.26**	**7.67**	**6.36**	**8.61**	9.13
12	**9.63**	**5.83**	**6.31**	**7.14**	**10.46**	**9.65**	**5.94**	**11.47**	11.57
13	**14.51**	22.37	16.62	**14.83**	**14.72**	**14.51**	22.10	16.97	15.15
14	**4.76**	8.68	**4.27**	**5.49**	**4.73**	**4.75**	8.58	**4.60**	6.38
15	**3.01**	8.48	**4.47**	6.46	**3.32**	**3.01**	8.38	**2.82**	4.58
16	**3.99**	7.71	**4.28**	6.82	**3.90**	**3.98**	7.69	**3.66**	4.64
17	**3.08**	6.03	**3.71**	6.88	**2.88**	**3.08**	6.01	**2.93**	3.73
18	**3.42**	4.95	**2.67**	5.19	**3.67**	**3.43**	4.87	**2.97**	4.81
19	**4.58**	**4.96**	**3.19**	**4.67**	**5.23**	**4.60**	**4.82**	**4.27**	7.22
20	**2.77**	**3.66**	**1.96**	**3.53**	**3.29**	**2.77**	**3.60**	**2.11**	4.96
21	**2.54**	**3.20**	**1.68**	**3.60**	**2.98**	**2.55**	**3.06**	**1.95**	4.43
22	**3.50**	**3.92**	**2.84**	**3.68**	**3.87**	**3.50**	**3.94**	**3.09**	5.42
23	**3.38**	**3.00**	**2.10**	**3.07**	**4.25**	**3.40**	**3.06**	**2.87**	6.69
24	**6.69**	**7.59**	**7.47**	8.20	**6.80**	**6.69**	**7.71**	**6.97**	7.72
25	**16.41**	18.13	18.25	17.73	16.78	**16.41**	18.07	17.29	16.56
26	**16.39**	**16.00**	**16.33**	**15.45**	**16.28**	**16.39**	**16.12**	16.57	16.45
27	**14.15**	**15.06**	**12.96**	**13.01**	**14.26**	**14.14**	**15.02**	**12.52**	15.28
28	**12.30**	**11.99**	**12.04**	**11.68**	**12.80**	**12.30**	**12.15**	**11.68**	13.52
29	**9.18**	**9.70**	**8.74**	**8.77**	**9.77**	**9.19**	**9.47**	**9.41**	11.54
30	**6.11**	**6.91**	**6.41**	**6.23**	**6.62**	**6.10**	**6.73**	**6.30**	7.03
				Summary statistics					
$\overline{\text{WMAE}}$	**8.25**	9.58	**7.80**	**8.58**	**8.57**	**8.26**	9.54	**8.50**	9.42
	(5.99)	(6.52)	(6.18)	(5.88)	(6.11)	(5.99)	(6.49)	(6.54)	(5.97)
# better than AR	27	17	26	19	27	27	17	22	–
# better than BI	15	8	21	14	14	15	8	–	–
# best	1	2	13	5	1	4	0	4	–
m.d.f.b.	**1.07**	2.40	**0.62**	**1.39**	**1.39**	**1.07**	2.36	**1.31**	2.24

With respect to the comparison between FA and BI approaches, neither presents clear dominance. Overall, simple averaging, LAD and IRMSE seem to outperform the BI approach with respect to the considered criteria. Results for PW and CLS are of similar quality as for the BI benchmark, while BI clearly provides better results than OLS and BMA for the considered time period. As mentioned above, these two approaches provide results that are also worse than the considered benchmark ARX model. As indicated by Figure 1, the November 2009 - June 2010 forecasting period also coincides with a more volatile and spiky behavior of spot electricity prices in Scandinavia. In particular during weeks 3, 6, 9 and 13, a number of price spikes is observed, while weeks 25-28 are characterized by a significantly higher volatility than prior weeks. On the other hand, results in Table 2 for these periods do not indicate that the benchmark approaches ARX and BI perform consistently better than the considered forecast

Table 3: *EEX 2009-2011 dataset*. WMAE for the seven forecast averaging schemes, one model selection technique (BI) and the benchmark model (AR). Summary statistics are provided in the bottom rows: $\overline{\text{WMAE}}$ is the mean value of WMAE for a given model (with standard deviation in parentheses), # *better* is the number of weeks a given averaging method is better than the benchmark or the BI, # *best* is the number of weeks a given averaging method performs best in terms of WMAE (i.e. excluding the individual methods), and finally *m.d.f.b.* is the mean deviation from the best model (averaging or individual) in each week. Emphasized in bold are the WMAE, $\overline{\text{WMAE}}$ and m.d.f.b. values lower than those of the benchmark. The out-of-sample period ranges from August 2, 2010 - February 27, 2011.

Week	Simple	OLS	LAD	PW	CLS	IRMSE	BMA	BI	AR
1	**7.04**	10.19	9.51	9.05	**7.01**	**6.95**	10.07	7.15	7.09
2	**7.69**	**7.32**	**7.08**	**7.51**	**7.52**	**7.69**	**7.52**	**8.10**	8.14
3	10.18	**9.68**	**9.58**	**9.76**	**9.82**	10.00	**9.86**	**9.97**	9.99
4	15.42	**14.84**	15.14	**14.81**	15.38	15.24	**14.89**	15.03	14.96
5	**10.41**	**9.45**	**9.40**	**9.58**	**10.92**	**10.56**	**9.39**	11.10	10.95
6	**10.85**	**10.88**	**10.65**	10.98	**10.87**	**10.87**	**10.88**	11.20	10.94
7	9.71	**8.62**	**8.87**	**8.92**	9.69	9.57	**8.73**	9.47	8.93
8	**11.64**	**10.52**	**10.16**	**10.69**	**12.14**	**11.94**	**10.44**	12.74	12.58
9	**10.71**	**8.18**	**7.70**	**8.25**	**10.04**	**10.69**	**8.15**	**11.21**	11.37
10	**10.71**	**9.48**	**9.35**	**9.98**	**10.47**	**10.82**	**9.52**	**10.36**	11.32
11	**12.23**	**8.93**	**8.72**	**8.97**	**11.24**	**12.19**	**8.85**	**10.33**	12.74
12	10.97	**9.62**	**9.51**	**10.01**	**10.35**	**10.81**	**9.63**	**10.09**	10.94
13	11.73	**9.93**	**9.69**	**10.67**	11.90	11.70	**9.92**	12.07	11.68
14	**13.34**	**13.22**	**13.09**	**13.39**	**13.56**	**13.51**	**13.23**	**13.88**	13.89
15	13.97	14.45	14.89	14.04	**13.55**	13.86	14.41	**13.35**	13.79
16	**10.57**	**6.98**	**6.86**	**7.02**	**9.20**	**10.62**	**6.96**	**8.61**	11.34
17	**10.73**	**8.25**	**7.95**	**8.56**	**9.59**	**10.67**	**8.29**	**9.31**	11.11
18	**14.87**	**12.26**	**12.27**	**12.87**	**13.69**	**14.76**	**12.35**	**13.39**	15.12
19	**19.09**	**17.75**	**18.06**	**17.91**	**17.40**	**18.90**	**17.85**	**17.02**	19.11
20	**16.45**	**13.25**	**13.75**	**13.11**	**15.15**	**16.35**	**13.38**	**15.04**	16.69
21	17.64	**17.40**	**17.19**	**17.58**	**17.39**	**17.63**	**17.43**	**17.37**	17.64
22	**17.32**	**17.02**	**16.45**	18.04	17.73	**17.34**	**17.08**	17.71	17.59
23	13.17	**12.26**	**12.39**	**12.44**	13.05	13.04	**12.36**	13.04	12.81
24	11.69	12.02	12.22	11.68	11.61	11.59	12.01	11.67	11.47
25	**10.56**	**6.60**	**6.25**	**6.75**	**8.42**	**10.57**	**6.50**	**8.16**	11.65
26	**8.21**	**6.26**	**5.92**	**6.82**	**7.59**	**8.17**	**6.18**	**7.60**	8.86
27	**18.29**	18.50	**18.26**	18.94	**16.98**	**18.21**	**18.40**	**17.00**	18.42
28	**12.33**	**10.69**	**10.32**	**10.85**	**11.56**	**12.44**	**10.67**	**11.51**	13.59
29	**9.47**	**7.12**	**6.79**	**7.54**	**8.38**	**9.47**	**7.17**	**8.34**	10.36
30	**8.32**	**4.83**	**4.51**	**5.15**	**6.39**	**8.27**	**4.85**	**6.31**	9.20
				Summary statistics					
$\overline{\text{WMAE}}$	**12.18**	**10.88**	**10.75**	**11.06**	**11.62**	**12.15**	**10.90**	**11.60**	12.48
	(3.20)	(3.62)	(3.73)	(3.64)	(3.21)	(3.18)	(3.63)	(3.18)	(3.10)
# better than AR	21	26	26	24	24	23	27	20	–
# better than BI	9	24	24	23	12	11	24	–	–
# best	0	3	19	2	1	2	1	2	–
m.d.f.b.	**1.73**	**0.43**	**0.30**	**0.61**	**1.17**	**1.70**	**0.45**	**1.15**	2.09

averaging techniques. However, for week 3 and week 13, the OLS and BMA approach yield results that are much worse than those of all other approaches what explains their overall inferior performance. Note especially the sharp increase in WMAE from week 12 to week 13. For week 12, these models achieved excellent results (almost halving the error of the ARX and BI approach) and might have been considered for continued use, underperforming only a week later when volatility picked up. We will discuss this instability of the approaches later on.

Table 3 provides the out-of-sample forecasting results for the EEX market for the time period August 2, 2010 - February 27, 2011. Recall that for this market no additional fundamental variable is used in the individual models so that the benchmark is actually an AR model. We find that, similar to the Nord pool market, also for the EEX all FA techniques provide significantly better results than the benchmark model. The different FA approaches outperform

the AR benchmark with respect to the weekly WMAE loss function between 65% and 90% of the time. In particular three of the combined forecasting approaches seem to perform quite well. The OLS, LAD and BMA method reduce the average WMAE by at least 12%, while m.d.f.b. is reduced by 78%-85% compared to the ARX model. Among these three models, it is again the LAD regression yielding the best results with respect to WMAE and m.d.f.b. LAD also provides the best performance of all models for 19 out of 30 weeks. The LAD emerges as a useful averaging scheme. Note that the absolute loss function we consider for evaluation of the forecasts matches the loss minimized by LAD. This, however, should not diminish the method's good performance since the weights are determined beforehand, while the evaluation is done ex post, and so it is possible to obtain far less satisfactory results.

Like for the Nord Pool market, also for the EEX dataset BI is substantially better than the benchmark AR model. Yet it is clearly outperformed by four of the combined forecasting techniques, namely OLS, LAD, PW and BMA. Also, BI is the best model for merely 2 out of the 30 weeks. On the other hand, simple averaging and IRMSE provide results that are worse than the best individual ex ante model on average. These models are outperformed by the BI model for over 60% of the time and also provide a higher average WMAE or m.d.f.b.

Overall, also for the EEX our results with respect to the superior performance of FA in comparison to the AR benchmark model are confirmed. All of the considered techniques provide better results for the considered loss functions and performance measures. However, only four out of seven FA techniques are able to outperform the BI model selection approach, while two of the methods yield results that are worse with respect to the considered measures.

Finally, we examine the results for the PJM market, where the out-of-sample period contains 30 weeks from June 19, 2011 to January 14, 2012. Interestingly, for this market, only four of the averaging techniques, namely simple averaging, LAD, CLS and IRMSE, are able to outperform the ARX model. These techniques provide better results than the ARX benchmark for approximately 60%-70% of the considered weeks and also yield slightly lower average WMAE and m.d.f.b. when the entire period is considered. However, for the PJM market, even the best performing FA techniques reduce the average WMAE by a small magnitude only of less than 3%, in comparison to the ARX model. LAD provides the best results for 12 out of 30 weeks, but the lowest average WMAE and m.d.f.b. are observed for simple averaging and IRMSE. Simple averaging also yields the best results for 9 out of 30 weeks. Note that the other three FA techniques, OLS, PW and BMA, yield considerably higher forecast errors than the ARX model, increasing the average WMAE by up to 25%. Also the BI selection approach performs only slightly better than the ARX benchmark. Note that for the PJM market, the examined out-of-sample period exhibits a times of quite volatile price behavior during week 4 and 5. However, the forecasting performance of the models does not seem to be dominated by the results for these weeks, since there are no substantial differences between the examined methods during this period. Overall, simple averaging, LAD, CLS and IRMSE provide the best results for the PJM market, while OLS, PW and BMA are outperformed by the benchmark ARX model and the BI approach.

5.2. Diebold-Mariano tests

In order to formally investigate the advantages from combining forecasts over selecting an individual model, we use the Diebold-Mariano (DM) test. Recall that predictions for all 24 hours of the next day are made at the same time using the same information set. Therefore, forecast errors for a particular day will typically exhibit high serial correlation as they are all affected by the same-day conditions. Therefore, we conduct these tests for each of the

Table 4: *PJM 2010-2012 dataset*. WMAE for the seven forecast averaging schemes, one model selection technique (BI) and the benchmark model (ARX). Summary statistics are provided in the bottom rows: $\overline{\text{WMAE}}$ is the mean value of WMAE for a given model (with standard deviation in parentheses), # *better* is the number of weeks a given averaging method is better than the benchmark or the BI, # *best* is the number of weeks a given averaging method performs best in terms of WMAE (i.e. excluding the individual methods), and finally *m.d.f.b.* is the mean deviation from the best model (averaging or individual) in each week. Emphasized in bold are the WMAE, $\overline{\text{WMAE}}$ and m.d.f.b. values lower than those of the benchmark. The out-of-sample period ranges from June 19, 2011 to January 14, 2012.

Week	Simple	OLS	LAD	PW	CLS	IRMSE	BMA	BI	ARX
1	12.95	23.90	16.04	22.08	15.13	13.02	23.07	15.64	12.14
2	**6.99**	17.31	8.30	16.57	**7.44**	**6.95**	16.82	7.74	7.56
3	**12.51**	14.71	**13.00**	14.98	**12.51**	**12.53**	14.86	**12.54**	13.78
4	**21.11**	22.90	**21.12**	**20.92**	**19.90**	**21.08**	22.77	**19.62**	21.41
5	**28.42**	33.49	31.26	**26.99**	**27.09**	**28.44**	33.72	**27.83**	29.43
6	13.94	22.80	14.34	20.38	**13.53**	13.94	22.76	**12.74**	13.71
7	**8.52**	12.50	**8.57**	12.31	**8.50**	**8.49**	12.40	**9.23**	9.81
8	8.85	12.16	9.99	15.38	9.99	8.88	12.19	9.86	8.68
9	**8.09**	10.73	9.04	11.25	8.55	**8.11**	10.59	**8.32**	8.41
10	13.07	17.00	13.36	16.30	13.31	13.07	17.11	13.27	12.50
11	**7.65**	10.25	**8.28**	9.58	**8.05**	**7.65**	10.22	8.43	8.41
12	9.29	14.32	9.92	11.78	9.74	9.31	14.07	9.20	8.73
13	10.56	11.83	10.23	12.80	10.23	10.55	11.72	10.34	10.09
14	**6.59**	8.99	**6.88**	**7.95**	**6.66**	**6.60**	8.88	**7.09**	8.00
15	**6.89**	11.10	**6.57**	8.96	**6.65**	**6.87**	10.82	**6.80**	7.16
16	**6.11**	8.40	**6.58**	7.64	**6.57**	**6.12**	8.43	**6.56**	6.85
17	**8.80**	11.50	**8.84**	10.40	**8.85**	**8.79**	11.47	**8.96**	9.18
18	**5.58**	6.96	**4.97**	7.67	**5.43**	**5.57**	7.00	5.92	5.68
19	**8.51**	12.87	9.38	9.25	9.17	**8.53**	13.01	8.94	8.79
20	9.65	12.44	**9.26**	11.17	9.88	9.65	12.59	9.69	9.34
21	9.01	9.91	8.70	10.91	9.06	9.00	9.92	9.11	8.50
22	10.40	10.96	**10.06**	12.68	10.64	10.40	11.00	11.00	10.08
23	**10.86**	**11.29**	**10.01**	12.51	**10.42**	**10.86**	**11.33**	**10.45**	11.65
24	**9.20**	9.82	**9.20**	9.82	**9.07**	**9.19**	10.01	9.49	9.36
25	**6.68**	7.77	**6.35**	8.10	6.87	**6.69**	7.88	**6.79**	6.87
26	**9.50**	**8.76**	**8.60**	10.11	**9.17**	**9.50**	**8.93**	**9.12**	9.52
27	**9.23**	**9.81**	**7.77**	10.44	**8.76**	**9.22**	9.97	**9.01**	9.86
28	**11.46**	**10.87**	**10.47**	11.69	**10.85**	**11.44**	**11.00**	**11.09**	11.65
29	**18.88**	**18.42**	**17.94**	19.23	**18.29**	**18.86**	**18.35**	**18.44**	19.02
30	**9.27**	12.08	**9.83**	10.76	10.23	**9.30**	12.39	10.22	10.12
				Summary statistics					
$\overline{\text{WMAE}}$	**10.62**	13.53	**10.83**	13.02	**10.68**	**10.62**	13.51	**10.78**	10.88
	(4.80)	(5.80)	(5.21)	(4.80)	(4.50)	(4.81)	(5.75)	(4.54)	(4.84)
# better than AR	21	5	19	3	19	21	4	16	–
# better than BI	19	4	18	1	18	19	4	–	–
# best	9	0	12	1	1	4	0	3	–
m.d.f.b.	**0.66**	3.57	**0.87**	3.06	**0.73**	**0.66**	3.55	**0.82**	0.92

$h = 1, .., 24$ hourly time series separately, using squared error losses of the model forecast:

$$L(\varepsilon_{h,t}) = (\varepsilon_{h,t})^2 = \left(P_{h,t} - \hat{P}_{h,t}\right)^2, \qquad h = \{1, \ldots, 24\}. \tag{13}$$

Note that Bordignon et al. (2013) used a similar approach, i.e. performed DM tests independently for each of the five half-hourly load periods considered in their study. Further note that we conducted additional DM tests for the absolute loss function, however, results were qualitatively similar and are omitted here.

For each forecast averaging technique and each hour we calculate the loss differential series $d_t = L(\varepsilon_{FA,t}) - L(\varepsilon_{ARX,t})$. We then conduct the DM tests for significant differences with respect to the performance of the benchmark ARX model (AR model for the EEX) and the BI selection technique. Note that we perform one-sided DM tests with the null hypothesis $H_0 : E(d_t) \leq 0$,

i.e. we test whether the considered forecast averaging approach can significantly outperform the ARX benchmark model at the 5% significance level. Figure 3 provides a graphical representation of the DM test statistic for each hour and method for two of the considered markets, namely the NP99 and PJM dataset. Note that for each forecast averaging technique and market we conduct 24 tests. Therefore, the figures illustrate results for 24 separately conducted tests for each averaging method.

Let us first consider the results for the NP99 dataset that are provided in the upper three panels of Figure 3. Recall that for this market the out-of-sample period refers to a relatively quiet period of spot price behavior where hardly any price spikes are observed. We find that most of the considered FA techniques significantly outperform the benchmark ARX approach for a large number of hours. In particular, simple averaging and IRMSE provide significantly better forecasting results than the ARX model for all hours considered. Also CLS and PW provide significantly better forecasting results than the benchmark model for 18, respectively 21, hours. On the other hand, LAD, OLS and BMA as the worst performing methods still significantly outperform the ARX model for 10, respectively 6, hours. Note that also the BI model selection strategy significantly outperforms the benchmark model for 14 hours.

As indicated by the lower three panels in Figure 3, results for the PJM market are not as clear cut as for the Nord Pool market. While simple averaging and IRMSE seem to provide significantly better forecasts than ARX for several of the considered hours, results are not as convincing for the other methods. CLS still seems to perform better than the benchmark model, but for most of the hours the results are not significant. On the other hand, LAD, OLS, PW and BMA all seem to perform worse than the ARX model. In particular OLS, PW and BMA provide forecasts that are consistently worse than the ARX model. Note that while the BI model selection strategy is signficantly better than the ARX benchmark for some of the hours, the majority of conducted tests indicate no significant difference between these models.

Table 5 aggregates results with respect to the benchmark ARX (or AR) model from all datasets. The upper part of the table provides a summary of the results for one-sided DM tests with the null hypothesis $H_0 : E(d_t) \geq 0$ based on the loss differential series $d_t = L(\varepsilon_{FA,t}) - L(\varepsilon_{ARX,t})$. The lower part of Table 5 also includes complementary results for the reverse null hypothesis: $H_0 : E(d_t) \leq 0$, namely whether the ARX benchmark model could significantly outperform the forecast averaging technique.

Note that all tests are conducted at the 5% significance level such that a rejection of the null suggests a significantly better (respectively, worse in the lower part of the table) performance of forecast averaging. In total 96 test-statistics were computed (4 datasets times 24 tests for each). Results where a forecast averaging technique outperforms the benchmark for at least 25% of the tests or vice versa (i.e. 6 or more significant results out of 24 for an individual market; 24 or more significant results out of 96 possible for all 4 datasets) are highlighted in bold.

For the two Nord Pool datasets, in particular simple averaging, PW, CLS and IRMSE perform significantly better than the ARX model for a high fraction of the conducted tests. Note that simple averaging and IRMSE significantly outperform the ARX benchmark for all 24 hours for the NP99 dataset and for 6 out of 24 hours for the NP10 dataset. Also PW and CLS perform well, outperforming the ARX model for 21, respectively 18, hours for the NP99 dataset and for 7 hours for the NP10 dataset. These findings complement results from Raviv et al. (2013) who also document favorable performance of constrained least squares estimation for Nord Pool data. BI model selection also outperforms the ARX benchmark for more than 55% of the conducted tests for the NP99 dataset. In addition, as indicated in the lower part of Table 5, the ARX model does not outperform any of the applied FA techniques or the BI approach for

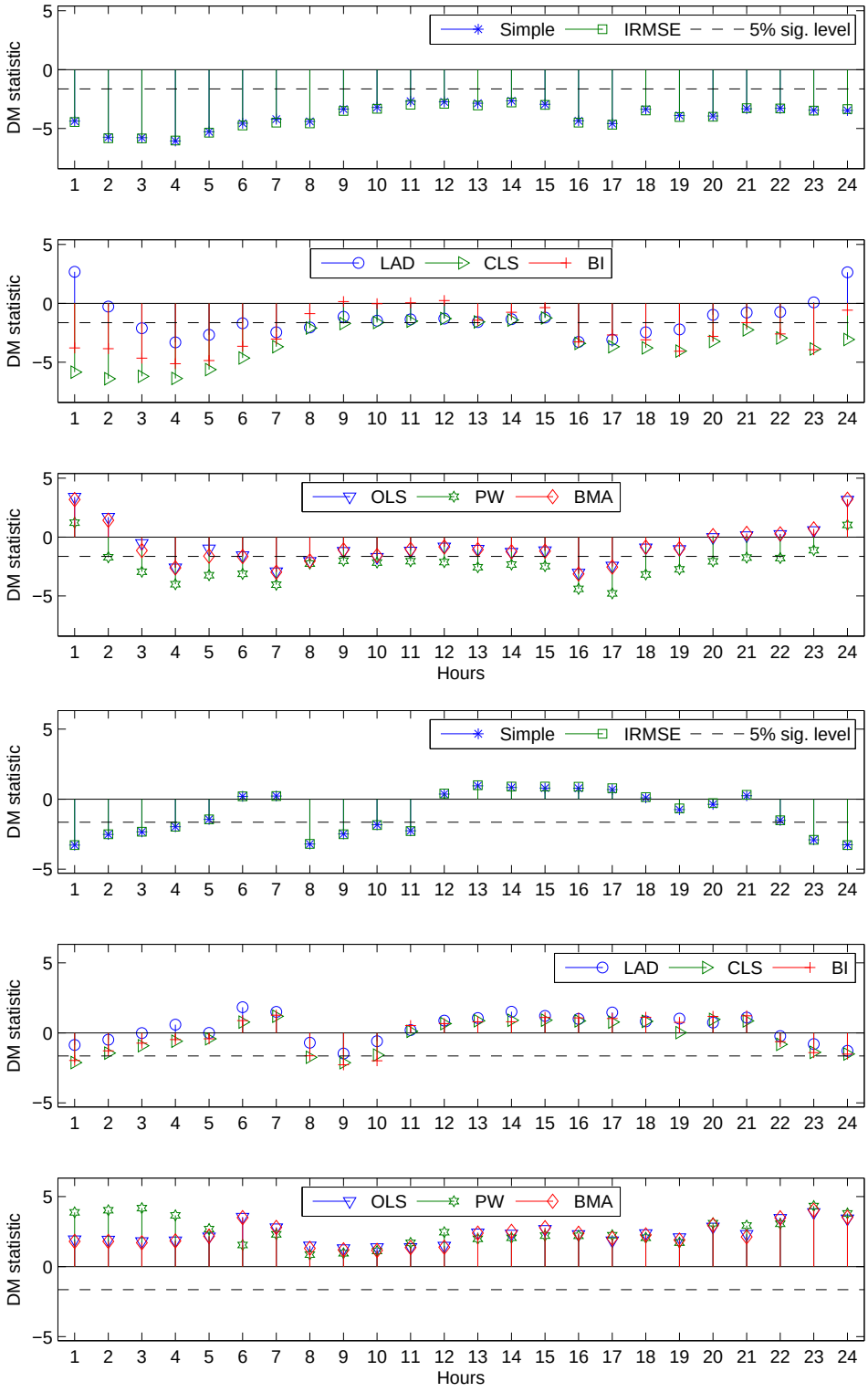

Figure 3: Results for conducted Diebold-Mariano tests for the NP99 (*upper three panels*) and PJM datasets (*lower three panels*). Tests are conducted separately for each of the 24 hours. The figures report the value of the test statistic for each test as well as the threshold where the null hypothesis $H_0 : E(d_t) \leq 0$ is rejected at the 5% significance level, where the loss function $d_t = L(\varepsilon_{FA,t}) - L(\varepsilon_{ARX,t})$.

Table 5: The percentage and the number (in parentheses) of hours for which an applied forecast averaging or model selection scheme is able to outperform the benchmark ARX (or AR for EEX) model (*upper part*) and for which the benchmark is able to outperform an averaging scheme (*lower part*), at the 5% significance level for the conducted Diebold-Mariano test. Note that the tests are applied separately for each of the 24 hourly time series in each market. The table also provides summary results for all markets, i.e. $4 \times 24 = 96$ conducted tests. Results where a technique outperforms the other for at least 25% of the series, i.e. at least six out of 24 hours for an individual market, respectively at least 24 out of 96 hours when all markets are considered, are highlighted in bold.

Method	NP99		NP10		EEX		PJM		Total	
Averaging better than ARX/AR										
Simple	**100%**	**(24)**	**25%**	**(6)**	**29%**	**(7)**	**42%**	**(10)**	**49%**	**(47)**
OLS	**25%**	**(6)**	17%	(4)	**75%**	**(18)**	0%	(0)	**29%**	**(28)**
LAD	**42%**	**(10)**	**29%**	**(7)**	**71%**	**(17)**	0%	(0)	**35%**	**(34)**
PW	**88%**	**(21)**	**29%**	**(7)**	**67%**	**(16)**	0%	(0)	**46%**	**(44)**
CLS	**75%**	**(18)**	**29%**	**(7)**	**50%**	**(12)**	13%	(3)	**42%**	**(40)**
IRMSE	**100%**	**(24)**	**25%**	**(6)**	**46%**	**(11)**	**42%**	**(10)**	**53%**	**(51)**
BMA	**25%**	**(6)**	17%	(4)	**79%**	**(19)**	0%	(0)	**30%**	**(29)**
BI	**58%**	**(14)**	4%	(1)	**46%**	**(11)**	13%	(3)	**30%**	**(29)**
ARX/AR better than averaging										
Simple	0%	(0)	4%	(1)	0%	(0)	0%	(0)	1%	(1)
OLS	13%	(3)	21%	(5)	0%	(0)	**79%**	**(19)**	**28%**	**(27)**
LAD	8%	(2)	4%	(1)	0%	(0)	4%	(1)	4%	(4)
PW	0%	(0)	21%	(5)	0%	(0)	**83%**	**(20)**	**26%**	**(25)**
CLS	0%	(0)	8%	(2)	0%	(0)	0%	(0)	2%	(2)
IRMSE	0%	(0)	4%	(1)	0%	(0)	0%	(0)	1%	(1)
BMA	8%	(2)	21%	(5)	0%	(0)	**79%**	**(19)**	**27%**	**(26)**
BI	0%	(0)	0%	(0)	0%	(0)	0%	(0)	0%	(0)

more than 25% of the hours for both datasets. Only for OLS, PW and BMA, the ARX model performs significantly better for 5 out of 24 hours for the NP10 dataset.

For the EEX, we find that apart from simple model averaging all FA techniques perform significantly better than the ARX model for at least 11 of the conducted tests. OLS, LAD, PW and BMA significantly outperform the benchmark model for even 67% or more of the considered hours. On the other hand, for the EEX market the ARX never provides significantly better results than any of the FA techniques. Also the BI selection technique significantly outperforms the ARX benchmark for 11 of the conducted tests, while ARX is never significantly better than BI.

Finally, for the PJM dataset, simple averaging and the IRMSE technique significantly outperform the ARX model for 10 out of 24 hours. It seems that for markets with few but very extreme periods of volatile price behavior these forecast averaging techniques perform best, what also confirms previous results on WMAE and m.d.f.b. However, none of the other methods, including BI, is able to outperform the ARX model for more than 3 out of 24 tests for the PJM market. On the other hand, as indicated in the lower part of the table, in particular OLS, PW and BMA are outperformed by the benchmark model for at least 19 out of 24 conducted tests. This confirms the poor performance of these methods for the more volatile PJM market.

The last column in Table 5 summarizes the results across all markets. Simple averaging and IRMSE outperform the benchmark ARX (or AR for EEX) model for more than 45 of the conducted tests, while PW and CLS are significantly better for 44, respectively 40 of the conducted tests. Therefore, these methods provide superior results in comparison to using a single forecasting model only. Note that also the BI selection method yields significantly better results than the benchmark model for 29 out of 96 tests, while the worst performing FA techniques, OLS and BMA, still outperform the ARX model for at least 28 of the conducted

Table 6: The percentage and the number (in parentheses) of hours for which an applied forecast averaging scheme is able to outperform the BI model selection approach (*upper part*) and for which the BI approach is able to outperform a forecast averaging scheme (*lower part*), at the 5% significance level for the conducted Diebold-Mariano test. Note that the tests are applied separately for each of the 24 hourly time series in each market. The table also provides summary results for all markets, i.e. $4 \times 24 = 96$ conducted tests. Results where a technique outperforms the other for at least 25% of the series, i.e. at least than six out of 24 hours for an individual market, respectively at least 24 out of 96 hours when all markets are considered, are highlighted in bold.

Method	NP99		NP10		EEX		PJM		Total	
Averaging better than BI model selection										
Simple	**67%**	**(16)**	**38%**	**(9)**	0%	(0)	13%	(3)	**29%**	**(28)**
OLS	4%	(1)	8%	(2)	**50%**	**(12)**	0%	(0)	16%	(15)
LAD	17%	(4)	**29%**	**(7)**	**54%**	**(13)**	0%	(0)	**25%**	**(24)**
PW	**50%**	**(12)**	**50%**	**(12)**	**33%**	**(8)**	0%	(0)	**33%**	**(32)**
CLS	**71%**	**(17)**	**29%**	**(7)**	0%	(0)	13%	(3)	**28%**	**(27)**
IRMSE	**58%**	**(14)**	**38%**	**(9)**	0%	(0)	13%	(3)	**27%**	**(26)**
BMA	0%	(0)	8%	(2)	**50%**	**(12)**	0%	(0)	15%	(14)
BI model selection better than averaging										
Simple	0%	(0)	0%	(0)	**38%**	**(9)**	0%	(0)	9%	(9)
OLS	**29%**	**(7)**	21%	(5)	0%	(0)	**75%**	**(18)**	**31%**	**(30)**
LAD	21%	(5)	4%	(1)	0%	(0)	13%	(3)	9%	(9)
PW	8%	(2)	8%	(2)	0%	(0)	**58%**	**(14)**	19%	(18)
CLS	0%	(0)	0%	(0)	0%	(0)	0%	(0)	0%	(0)
IRMSE	0%	(0)	0%	(0)	**33%**	**(8)**	0%	(0)	8%	(8)
BMA	**29%**	**(7)**	21%	(5)	0%	(0)	**71%**	**(17)**	**30%**	**(29)**

tests. However, the latter two techniques, as well as PW, are also outperformed by the ARX technique for a relatively high number of tests (between 25 and 27). Notwithstanding, all other FA techniques and the BI method show significantly worse results than the ARX benchmark only for a very small number of hours, i.e. less than 5%. Overall, our results for conducted DM tests confirm results previously reported for the WMAE and m.d.f.b. measures and strongly support the superior performance of most FA techniques in comparison to the ARX benchmark model. Also the BI selection technique tends to perform significantly better than the ARX model. Only the OLS, PW and BMA approaches are also outperformed by the benchmark ARX model for a relatively high number of tests. We discuss possible reasons for this later on.

In a last step we also conduct DM tests in order to compare the FA techniques with the BI approach, see Table 6. Recall that BI could be viewed as a special case of forecast averaging with degenerate weights as indicated by equation (11). However, it is essentially a model (or forecast) selection scheme and may be considered as a second, more sophisticated, benchmark against the choice of forecast averaging. As pointed out before, similar to the FA techniques, BI generally performed significantly better than the ARX benchmark model. We now further continue to discuss whether results are comparable to the rest of the FA techniques. Considering the upper part of Table 6, we find that in particular simple averaging, PW, CLS and IRMSE seem to outperform BI for three of the considered datasets, namely the two Nord Pool and the EEX out-of-sample test periods. On the other hand, BI significantly outperforms OLS, PW, and BMA for the PJM dataset for at least 14 and up to 18 of the conducted tests (see lower part of the table). Overall, five out of seven FA methods provide significantly better results than the BI technique for at least 25% of the conducted tests. On the other hand, the BI selection scheme also seems to give a better performance than two of the FA techniques, namely the unrestricted averaging schemes OLS and BMA. Thus, while forecast averaging overall still seems to be the preferred technique, results are not as clear-cut as for the ARX benchmark model.

5.3. Summary

Tables 5 and 6 taken together along with the previous analysis are very much in line with the abundant related literature regarding forecast averaging. First, combining forecasts is superior in terms of accuracy to selecting forecasts. Second, though there is nothing sacred in restricting the weights during the averaging process, it is preferred over an unrestricted version. A restricted version such as CLS provides a sub-optimal solution in-sample, however, the unrestricted version which provides a global optimum in-sample, potentially provides extremely bad results out-of-sample, as is viewed for week 13 in Table 2. Third, next to CLS and IRMSE also simple forecast averaging emerges as a stable choice which may not be optimal, yet apart from the EEX data, our significance testing shows that one will never significantly worsen prediction accuracy by using an equally weighted combined forecast instead of selecting an ARX/AR model or the best individual ex ante model. This robustness property of the equal weighting scheme is shown, perhaps counter-intuitively, to be expected by Smith and Wallis (2009) when the quality of the individual forecasts is similar. This argument does not hold true when applied to the EEX dataset. We believe that the reason is the lower number of models we consider for this dataset. Due to data limitations, we only use 6 individual forecasts, instead of the usual 12 for the other datasets. This means that there is less noise in parameter estimation, so in the case of EEX, a more sophisticated averaging method might be a better choice. Support for this argument is found in the portfolio management literature; DeMiguel et al. (2009) show that optimal portfolio construction is more likely to outperform a simple weighting scheme when fewer assets are present, so fewer weights to estimate. On the other hand, the equally weighted portfolio is expected to outperform optimal weighting in the presence of substantial estimation noise.

6. Conclusions

We examine possible accuracy gains from forecast averaging in the context of electricity spot price prediction. While there is a significant number of studies on the use of forecast combinations for predicting economic and financial variables, there is only a small number of applications of these important techniques in the area of electricity markets, and even fewer where electricity spot price forecasts are discussed. Our paper can be considered as an extension of the empirical study by Bordignon et al. (2013). While our findings are similar in spirit, they are not as clear-cut, possibly due to the fact that we consider more (12 vs. 5) and different individual models, more datasets (4 vs. 1) and, most importantly, more diverse averaging schemes.

Namely, we apply seven averaging approaches and one selection method and perform a backtesting analysis on electricity spot (or day-ahead) prices for the Scandinavian Nord Pool market, the European Energy Exchange in Germany and the Pennsylvania-New Jersey-Maryland Interconnection (PJM) in the US. We compare the averaging techniques in a realistic setting where market participants have to decide ex ante which individual model will be applied for forecasting. We evaluate the results using two feasible benchmarks:

- an ARX model (or an AR model for the EEX dataset), based on its ease of implementation and good forecasting performance in other studies (see e.g. Conejo et al., 2005; Gonzalez et al., 2012; Misiorek et al., 2006; Nogales et al., 2002; Raviv et al., 2013), and
- the best individual (BI) ex ante model, i.e. at each point in time t, pick the model that performed best up to time $t - 24$.

Note that the latter is essentially a model (or forecast) selection scheme, but we can also view it as a special case of forecast averaging with degenerate weights.

Overall, our findings support the additional benefits of combining forecasts for deriving more accurate price forecasts in the considered markets. Five out of seven forecast averaging methods clearly outperform the benchmark ARX model. Also, the majority of averaging techniques seem to outperform the best individual ex ante scheme. That said, methods that allow for unconstrained weights, such as OLS averaging should be avoided. Due to the specific behavior of electricity prices an unconstrained weight vector which, if poorly estimated, can be severely punished in terms of accuracy, arguably much more than for other commodities.

Despite extensive research efforts, in general, and specifically for electricity markets, there is no clear prior guidance as to which forecast combination scheme works best. A point apparent from our results as well. IRMSE and simple averaging perform best with respect to the ARX model – they are significantly more accurate than the benchmark in about 50% of cases and significantly less accurate only in 1% of cases. While these combination schemes still perform extremely well against the best individual ex ante model in general, for the German EEX market they are significantly less accurate than BI in as many as 33-38% of cases. On the other hand, CLS averaging stands out as a choice which may not be optimal, but will never significantly worsen prediction accuracy compared to the best individual ex ante model. As no single forecasting method clearly dominates all others for all datasets considered, we recommend a backtesting exercise to identify the preferred forecast averaging method for the data at hand.

Acknowledgements

This paper has benefited from conversations with the participants of the 66th European Meeting of the Econometric Society, the Conference on Energy Finance EF2012, the Energy Finance Christmas Workshop (EFC12), the European Energy Market (EEM13) Conference, the SIRE Conference on Finance and Commodities and the seminars at Macquarie University, National University of Singapore and Wrocław University of Technology. This work was supported by funds from the Australian Research Council through grant no. DP1096326 and the National Science Centre (NCN, Poland) through grant no. 2011/01/B/HS4/01077.

Bibliography

Aksu C., Sevket I.G. (1992) An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts. International Journal of Forecasting 8(1), 27-43.

Andrawis, R., Atiya, A., Saavedra, A. (2011) Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of Forecasting 27(3), 870-886.

Ball, C.A., and W.N. Torous (1983) A simplified jump process for common stock returns. Journal of Finance and Quantitative Analysis 18(1), 53-65.

Bates, J. M., Granger, C. W. (1969), The combination of forecasts, Operations Research Quarterly, 20(4), 451-468.

Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S. (2007) Spot and derivative pricing in the EEX power market. Journal of Banking and Finance 31, 3462-3485.

Bordignon, S., Bunn, D. W., Lisi, F., Nan, F. (2013), Combining day-ahead forecasts for British electricity prices. Energy Economics 35, 88-103.

Bunn, D.W. (2000) Forecasting loads and prices in competitive power markets. Proceedings of the IEEE 88(2), 163-169.

Bunn, DW., ed. (2004) Modelling Prices in Competitive Electricity Markets. Wiley, Chichester.

Bunn, D. W., Andresen, A., Chen, D., Westgaard, S. (2013) Analysis and forecasting of electricity price risks with quantile factor models. Working Paper.

Cao, R., Hart, J.D., Saavedra, A. (2003) Nonparametric maximum likelihood estimators for AR and MA time series. Journal of Statistical Computation and Simulation 73(5), 347-360.

Clemen, R.T. (1989) Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559-583.
Conejo, A.J., Contreras, J., Espinola, R., Plazas, M.A. (2005) Forecasting electricity prices for a day-ahead pool-based electric energy market. International Journal of Forecasting 21(3), 435-462.
Crane, D.B., Crotty, J.R. (1967) A two-stage forecasting model: Exponential smoothing and multiple regression. Management Science 6(13), B501-B507.
DeMiguel, V., Garlappi, L., Uppal, R. (2009) Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? Review of Financial Studies 22(5), 1915-1953.
Diebold F., Pauly P. (1987) Structural change and the combination of forecasts, Journal of Forecasting 6, 21-40
Eydeland, A., Wolyniec, K. (2013) Energy and Power Risk Management (2nd ed.). Wiley, Hoboken, NJ.
Genre, V., Kenny, G., Meyler, A., Timmermann, A. (2013) Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting 29(1), 108-121.
Gonzalez, V., Contreras, J., Bunn, D.W. (2012) Forecasting power prices using a hybrid fundamental-econometric model. IEEE Transactions on Power Systems 27(1), 363-372.
Granger, C.W., Ramanathan, R. (1984) Improved methods of combining forecasts. Journal of Forecasting 3, 197-204.
Hsieh, D.A., and C.F. Manski (1987) Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. Annals of Statistics 15, 541-551.
Huisman, R. (2009) An Introduction to Models for the Energy Markets. Risk Books.
Huisman, R., Huurman, C., Mahieu, R. (2007) Hourly electricity prices in day-ahead markets. Energy Economics 29, 240-248.
Janczura, J., Trueck, S., Weron, R., Wolff, R. (2013) Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling, Energy Economics 38, 96-110.
Jonsson, T., Pinson, P., Madsen, H., Nielsen, H.A. (2013) Predictive densities for day-ahead electricity prices using time-adaptive quantile regression. Applied Energy, submitted.
Koenker R. (2005) Quantile Regression. Cambridge University Press.
Løland, A., Ferkingstad, E., Wilhelmsen, M. (2012) Forecasting transmission congestion. Journal of Energy Markets 5(3), 65-83.
Madigan, D., Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 89, 1535-1546.
de Menezes, L.M., Bunn, D.W., Taylor, L.W. (2000) Review of guidelines for the use of combined forecasts. European Journal of Operations Research 120, 190-204.
Misiorek, A., Trück, S., Weron, R. (2006) Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models. Studies in Nonlinear Dynamics and Econometrics 10(3), Article 2.
Newbold, P., Granger, C.W. (1974) Experience with forecasting univariate time series and the combination of forecasts, Journal of the Royal Statistical Society A, 137, 131-164.
Nogales, F.J., Contreras, J., Conejo, A.J., Espinola, R. (2002) Forecasting next-day electricity prices by time series models. IEEE Transactions on Power Systems 17, 342-348.
Nowotarski, J., Tomczyk, J., Weron, R. (2013) Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. Energy Economics 39, 13-27.
Peña, J.I. (2012) A note on panel hourly electricity prices. Journal of Energy Markets 5(4), 81-97.
Raftery, A.E., Painter, I.S., Volinsky, C.T. (2005) BMA: An R package for Bayesian Model Averaging. R News 5(2), 2-8.
Raviv, E., Bouwman, K.E., van Dijk, D. (2013) Forecasting day-ahead electricity prices: Utilizing hourly prices. Tinbergen Institute Discussion Paper 13-068/III. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2266312.
Sevket I.G., Aksu C. (1989) N step combinations of forecasts. Journal of Forecasting 8(3), 253-267.
Shahidehpour, M., Yamin, H., Li, Z. (2002) Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management. Wiley.
Smith, D.G. (1989) Combination of forecasts in electricity demand prediction, Journal Of Forecasting, 8, 349-356.
Smith, J., Wallis, K. (2009) A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71(3), 331-355.
Stock, J.H., Watson, M.W. (2004) Combination Forecasts of Output Growth in a Seven-country Data Set, Journal of Forecasting 23, 405-430.
Taylor, J.W., Majithia, S. (2000) Using combined forecasts with changing weights for electricity demand profiling. Journal of the Operational Research Society 51, 72-82.
Taylor, J.W. (2010) Triple seasonal methods for short-term electricity demand forecasting. European Journal of

Operations Research 204, 139-152.
Timmermann, A.G. (2006) Forecast combinations. In: Elliott, G., Granger, C.W., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Elsevier, 135-196.
Tong, H., and K.S. Lim (1980). Threshold autoregression, limit cycles and cyclical data, Journal of the Royal Statistical Society B 42, 245-292.
Weron, R. (2006) Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Wiley, Chichester.
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

HSC Research Report Series 2013

For a complete list please visit http://ideas.repec.org/s/wuu/wpaper.html

01 *Forecasting of daily electricity spot prices by incorporating intra-day relationships: Evidence form the UK power market* by Katarzyna Maciejowska and Rafał Weron

02 *Modeling and forecasting of the long-term seasonal component of the EEX and Nord Pool spot prices* by Jakub Nowotarski, Jakub Tomczyk and Rafał Weron

03 *A review of optimization methods for evaluation of placement of distributed generation into distribution networks* by Anna Kowalska-Pyzalska

04 *Diffusion of innovation within an agent-based model: Spinsons, independence and advertising* by Piotr Przybyła, Katarzyna Sznajd-Weron and Rafał Weron

05 *Going green: Agent-based modeling of the diffusion of dynamic electricity tariffs* by Anna Kowalska-Pyzalska, Katarzyna Maciejowska, Katarzyna Sznajd-Weron and Rafał Weron

06 *Relationship between spot and futures prices in electricity markets: Pitfalls of regression analysis* by Michał Zator

07 *An empirical comparison of alternate schemes for combining electricity spot price forecasts* by Jakub Nowotarski, Eran Raviv, Stefan Trueck and Rafał Weron