A Note on Averaging Day-Ahead Electricity Price Forecasts Across Calibration Windows

Katarzyna Hubicka, Grzegorz Marcjasz, and Rafał Weron

***Abstract*—We propose a novel concept in energy forecasting and show that averaging day-ahead electricity price forecasts of a predictive model across 28–728 day calibration windows yields better results than selecting only one "optimal" window length. Even more significant accuracy gains can be achieved by averaging over a few, carefully selected windows.**

***Index Terms*—Electricity price forecasting, combining forecasts, calibration window, autoregression, NARX neural network, committee machine, Diebold–Mariano test.**

I. Introduction

Most day-ahead *electricity price forecasting* (EPF) studies focus on developing model structures that better represent the temporal and inter-variable dependencies, feature (i.e., input variable) selection, implementing faster and more efficient estimation algorithms or finding optimal weights for combined forecasts [1]. However, very few studies in energy forecasting and – to our best knowledge – none in EPF try to find the optimal length of the calibration window or average forecasts obtained from models estimated on different windows. Here we address this important, but overlooked issue.

In the econometric literature, some researchers argue that forecasting performance is sensitive to the choice of the calibration window and in the presence of structural breaks (i.e., abrupt and unexpected changes in the underlying process) it may be better to combine forecasts based on windows of different lengths [2]. Longer windows allow for more precise estimation of model parameters, but shorter better adapt to changes. Hence, forecasts obtained from different windows will address distinct features of the underlying process.

While some authors try to develop optimal weighting schemes [3], the general conclusion from these studies is that a simple arithmetic average across all windows is robust and hard to outperform; an outcome that reminds of the results obtained when combining forecasts from different models [4] or from models calibrated to different data subsets [5].

We illustrate our concept of averaging forecasts across calibration windows using two popular models: an autoregression with exogenous variables (ARX) and its non-linear counterpart – a NARX neural network. We provide evidence that averaging forecasts across all window lengths ranging from 28 to 728 days outperforms selecting (even *ex-post*) only one 'optimal' window. Furthermore, we argue that in the context of electricity markets, where the time series of interest (prices, loads) are characterized by weekly and annual seasonal behavior, there may yet be a better alternative. Indeed, as we show below, averaging across a few short (e.g., 28, 56 and 84 days) and a few long (e.g., 714, 721 and 728 days) window lengths brings further, significant accuracy gains.

II. Methodology

Like in many EPF studies, the modeling is implemented here within a 'multivariate' framework [6]. We explicitly use a 'day × hour', matrix-like structure with $p_{d,h} \equiv \log(P_{d,h})$ representing the electricity log-price for day d and hour h, and consider two models, each consisting of 24 submodels – one for each hour of the day. The **ARX** model is based on a well performing autoregressive structure [4], [6]–[8]: $p_{d,h} = \beta_{h,0} + \beta_{h,1} p_{d-1,h} + \beta_{h,2} p_{d-2,h} + \beta_{h,3} p_{d-7,h} + \beta_{h,4} p_{\min} + \beta_{h,5} z_t + \sum_{i \in \{1,6,7\}} \beta_{h,i+5} D_i + \varepsilon_{d,h}$, where $p_{\min} = \min_h \{p_{d-1,h}\}$ creates a link with all yesterday's prices, not just the prices for the same hour, z_t is (the logarithm of) the day-ahead load forecast, D_i is the dummy variable for day-of-the-week i and $\varepsilon_{d,h}$ is the noise term.

Like in [9], the **NARX** model is a recurrent neural network with the same inputs as the **ARX** model, one hidden layer consisting of 5 neurons (with tangent sigmoid transfer functions) and an output layer with one neuron yielding $p_{d,h}$ (with a linear transfer function). It is trained in Python using the incremental scheme of the *Fast Artificial Neural Networks* library [10]. Since the algorithm is initialized using a random starting point, different estimates and hence forecasts are obtained for each run (and each day and hour). To decrease the variance we consider a committee machine, i.e., repeat every training and forecasting exercise 5 times and average forecasts on an hour-by-hour basis across the runs, as in [9], [11].

As the test environment we consider a publicly available dataset from the Global Energy Forecasting Competition 2014

Manuscript received June 7, 2018; revised August 2, 2018; accepted September 6, 2018. Date of publication September 10, 2018; date of current version December 14, 2018. This work was supported by the National Science Center (NCN, Poland) under Grants 2016/23/G/ST1/04083 (to KH), 2016/23/G/HS4/01005 (to GM), and 2015/17/B/HS4/00334 (to RW). Paper no. PESL-00122-2018. *(Corresponding author: Rafał Weron.)*

K. Hubicka is with the Faculty of Pure & Applied Mathematics, Wrocław University of Science and Technology, Wrocław 50-370, Poland (e-mail: hhubickak@gmail.com).

G. Marcjasz and R. Weron are with the Faculty of Computer Science and Management, Department of Operations Research, Wrocław University of Science and Technology, Wrocław 50-370, Poland (e-mail: gelusz@hotmail.co.uk; rafal.weron@pwr.edu.pl).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSTE.2018.2869557

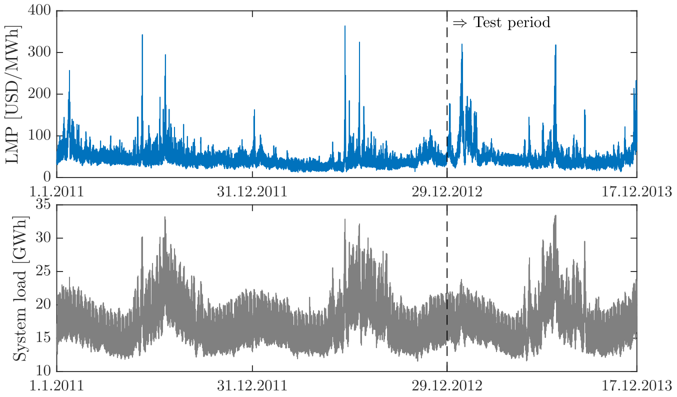

Fig. 1. GEFCom2014 hourly locational marginal prices (LMPs; *top*) and hourly day-ahead predictions of system load (*bottom*) for the period 1.1.2011–17.12.2013. The vertical dashed lines mark the beginning of the test period.

(GEFCom2014) [12]. It comprises hourly locational marginal prices (LMP) and day-ahead predictions of hourly system loads from the period 1.1.2011–17.12.2013, see Fig. 1. We use a rolling window scheme, with data spanning the most recent $T = 28, \ldots, 728$ days. The length of the largest window sets a limit on the size of the test sample: the model is initially calibrated to data from 1.01.2011–28.12.2012 and forecasts for all 24 h of 29.12.2012 are determined. Next, the windows are rolled forward by one day and forecasts for all 24 h of 30.12.2012 are computed, etc.

III. Results

We evaluate the forecasts in terms of the *Mean Absolute Error* over the full test period of $D = 354$ days (29.12.2012-17.12.2013): $\text{MAE} = \frac{1}{24D}\sum_{d=1}^{D}\sum_{h=1}^{24}|\widehat{\varepsilon}_{d,h}|$, where $\widehat{\varepsilon}_{d,h}$ is the prediction error for day d and hour h. In Fig. 2 we plot MAE errors as a function of the calibration window length (the same for all $D = 354$ days) and compare them with errors obtained by averaging electricity price forecasts across calibration windows of different lengths. We use **Win**(T) to denote the forecast for a calibration window of length T days and **AW**($\mathcal{T}$) to denote an average forecast across a set of windows. We use Matlab's notation for the latter, e.g., $\mathcal{T} = \{28, 728\}$ refers to 28- and 728-day windows and $\mathcal{T} = \{28{:}28{:}728\}$ to 26 windows: 28-, 56- ($=2 \times 28$), ..., 728-day.

MAE errors for selected window sets are reported in Table I and illustrated in Fig. 2. Clearly, the MAE for AW(28:728), i.e., an average across all windows, is lower than for any of the Win(T)'s. It is also 1.2–1.35% better than the longest window, see the '%chng.' columns in Table I; the relative changes are computed as $\log(\text{MAE}_{\text{Win}(728)}/\text{MAE}_{\text{AW}(\mathcal{T})})$. However, AW(28:728) is not computationally efficient, especially for the **NARX** model, which takes longer to train than a regression and requires a committee machine; note that 5 runs still yield volatile forecasts, see the scattered blue dots in Fig. 2. As a remedy we could select a few windows and average forecasts only for those few. But how to select them?

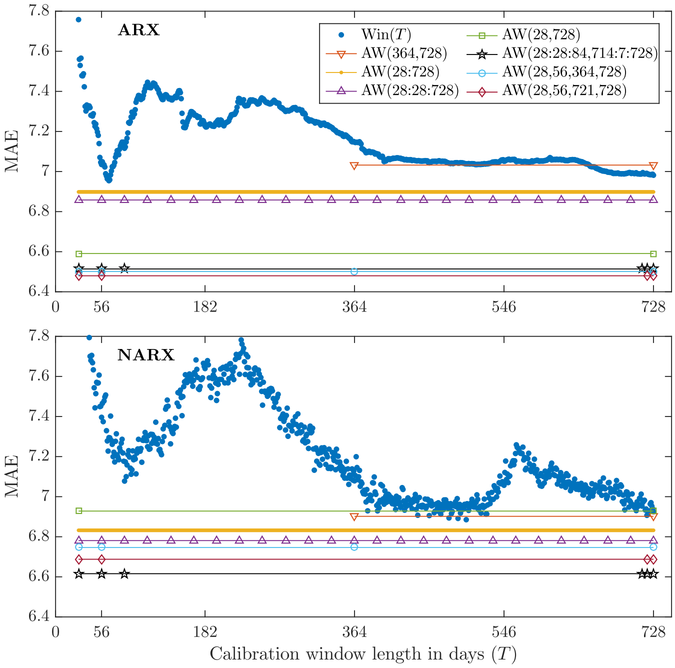

Fig. 2. Mean Absolute Errors (MAE) as a function of the window length $T = 28, \ldots, 728$ (blue circles) and obtained by averaging forecasts across calibration windows (lines with symbols representing window lengths) for the **ARX** (*top*) and **NARX** (*bottom*) models.

TABLE I
Mean Absolute Errors (MAE) for Selected Calibration Window Sets and the **ARX** (in Descending Order) and **NARX** Models. Relative Changes (%chng.) with Respect to the Win(728) MAE Within Each Model Class are also Reported

	ARX		**NARX**	
Windows	MAE	%chng.	MAE	%chng.
Win(28)	7.758	−10.5%	8.394	−19.2%
Win(364)	7.147	−2.35%	7.079	−2.20%
AW(364,728)	7.032	−0.72%	6.902	0.34%
Win(728)	6.982	—	6.925	—
AW(28:728)	6.898	1.20%	6.832	1.35%
AW(28:7:728)	6.891	1.30%	6.793	1.93%
AW(28:14:728)	6.879	1.48%	6.787	2.01%
AW(28:28:728)	6.858	1.78%	6.781	2.10%
AW(56,728)	6.638	5.05%	6.718	3.03%
AW(28,728)	6.591	5.76%	6.928	−0.04%
AW(28:28:84,714:7:728)	6.514	6.93%	**6.616**	**4.57%**
AW(28,56,728)	6.509	7.01%	6.888	0.54%
AW(28,56,364,728)	6.501	7.13%	6.746	2.61%
AW(28,56,721,728)	**6.480**	**7.46%**	6.688	3.49%

First, let us limit the search to windows whose length is a multiple of 7 days. Although the weekly seasonality does not seem to affect **ARX** forecasts, it is visible for **NARX**, especially for shorter windows. Second, having in mind that longer windows allow for modeling trends and shorter better adapt to changes, let us start with $\mathcal{T} = \{28, 728\}$ and expand the set by adding two windows at a time: one m (7 or 28) days longer than the longest of the short windows and one n (7 or 28) days shorter than the shortest of the long ones. E.g., if $m = 28$ and $n = 7$ then $\mathcal{T}$ is expanded to $\{28, 56, 721, 728\}$, $\{28{:}28{:}84, 714{:}7{:}728\}, \ldots,$ $\{28{:}28{:}560, 595{:}7{:}728\}$ and $\{28{:}28{:}560, 588, 595{:}7{:}728\}$. The results are plotted in Fig. 3. Clearly, the gains from averaging **ARX** forecasts are higher than for **NARX**. However, the main message is the same in both panels: the mixed combination ($m = 28$ and $n = 7$; blue circles) yields better forecasts than

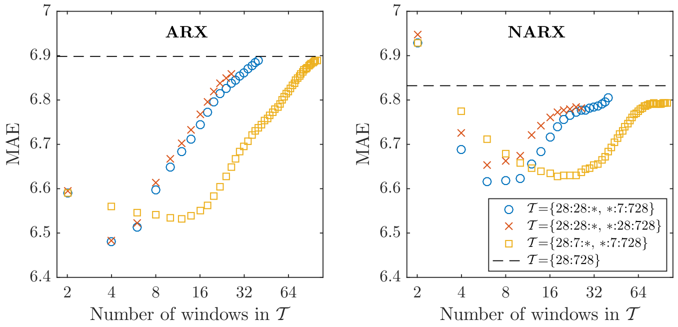

Fig. 3. Mean Absolute Errors (MAE) for the **ARX** (*left*) and **NARX** (*right*) models as a function of the number of windows in $\mathcal{T}$, for an iterative procedure described in the text. We use a wildcard ($*$) to denote the 'middle' windows in $\mathcal{T} = \{28{:}m{:}*, *{:}n{:}728\}$ and let $m, n = 7$ or 28.

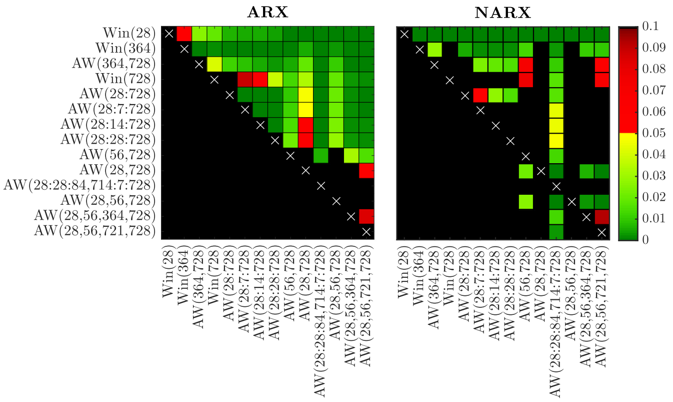

Fig. 4. Results of the multivariate DM test for selected window sets and the **ARX** (*left*) and **NARX** (*right*) models. We use a heat map to indicate the range of the p-values – the closer they are to zero (→ dark green) the more significant is the difference between the forecasts of a set on the X-axis (better) and the forecasts of a set on the Y-axis (worse).

the homogeneous ones ($m, n = 7$ or 28) and the lowest MAEs are obtained for as few as 4 (6 for **NARX**) windows. The latter result is particularly important in view of the much higher computational cost of using neural networks. Note also, that the best **ARX** window set performs very well for **NARX** and vice versa.

To provide statistically significant conclusions on the differences in forecasting performance we follow [13] and conduct the *multivariate* variant of the Diebold-Mariano (DM) test on a day-by-day basis, i.e., for each day a model's score is computed as $\sum_{h=1}^{24} |\widehat{\varepsilon}_{d,h}|$. As in the standard DM test, we assume that the loss differential series is covariance stationary. In Fig. 4 we plot the p-values of the conducted pairwise comparisons: green and yellow squares indicate statistical significance at the 5% level (with the darkest green corresponding to close to zero p-values), red squares indicate weak significance with p-value $\in [5\%, 10\%)$, while black denote no significance (p-value $\geq 10\%$). For instance, the first row in the right panel is green, so that the forecasts of the **NARX** model for every window set significantly outperform those for Win(28), while the column which corresponds to AW(28:28:84,714:7:728) in the same panel is green or yellow, meaning that this window set leads to significantly better forecasts than all other. Interestingly, in the left panel the best four models (the ordering is the same as in Table I) essentially do not differ in terms of significance. Hence, we recommend AW(28:28:84,714:7:728) as it leverages accurate predictions for both model classes with computational efficiency and is not significantly outperformed by any other window set.

IV. Conclusion

The extremely simple idea we advocate here may have far reaching consequences for electricity price forecasting and energy forecasting in general. Our empirical study shows that significant accuracy gains can be achieved by averaging day-ahead electricity price forecasts of a predictive model across different calibration windows, compared to the results obtained for one selected window length. Naturally, our study can and should be extended in several directions. In particular, to more advanced forecasting models (SVMs, LASSO estimated regressions, etc.), to other datasets (other markets, loads, wind power generation, etc.), and to more sophisticated weighting schemes [14]. Finally, the economic impact of the improved forecasts should be evaluated [15]. This, however, is left for future research.

References

[1] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *Int J. Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014.

[2] M. Pesaran and A. Timmermann, "Selection of estimation window in the presence of breaks," *J. Econometrics*, vol. 137, no. 1, pp. 134–161, 2007.

[3] J. Tian and H. Anderson, "Forecast combinations under structural break uncertainty," *Int. J. Forecasting*, vol. 30, no. 1, pp. 161–175, 2014.

[4] J. Nowotarski, E. Raviv, S. Trück, and R. Weron, "An empirical comparison of alternate schemes for combining electricity spot price forecasts," *Energy Econ*, vol. 46, pp. 395–412, 2014.

[5] J. Nowotarski, B. Liu, R. Weron, and T. Hong, "Improving short term load forecast accuracy via combining sister forecasts," *Energy*, vol. 98, pp. 40–49, 2016.

[6] F. Ziel and R. Weron, "Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks," *Energy Econ.*, vol. 70, pp. 396–420, 2018.

[7] P. Gaillard, Y. Goude, and R. Nedellec, "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting," *Int. J. Forecasting*, vol. 32, no. 3, pp. 1038–1050, 2016.

[8] F. Ziel, "Forecasting electricity spot prices using LASSO: On capturing the autoregressive intraday structure," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4977–4987, Nov. 2016.

[9] G. Marcjasz, B. Uniejewski, and R. Weron, "On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks," *Int. J. Forecasting*, to be published, doi: 10.1016/j.ijforecast.2017.11.009.

[10] S. Nissen, "Large Scale Reinforcement Learning using Q-SARSA(λ) and Cascading Neural Networks," M.Sc. Thesis, Dept. Comput. Sci., University of Copenhagen, Copenhagen, Denmark, 2007.

[11] N. Shrivastava and B. Panigrahi, "A hybrid wavelet-elm based short term price forecasting for electricity markets," *Int. J. Electric. Power Energy Syst.*, vol. 55, pp. 41–50, 2014.

[12] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," *Int. J. Forecasting*, vol. 32, no. 3, pp. 896–913, 2016.

[13] B. Uniejewski, R. Weron, and F. Ziel, "Variance stabilizing transformations for electricity spot price forecasting," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2219–2229, Mar. 2018.

[14] G. Marcjasz, T. Serafin, and R. Weron, "Selection of calibration windows for day-ahead electricity price forecasting," *Energies*, vol. 11, no. 9, p. 2364, 2018, doi: 10.3390/en11092364.

[15] A. Doostmohammadi, N. Amjady, and H. Zareipour, "Day-ahead financial loss/gain modeling and prediction for a generation company," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3360–3372, Sep. 2017.