A value-based evaluation methodology for renewable energy supply prediction

Robert Ulbricht[2], Bijay Neupane[3], Martin Hahmann[1], and Wolfgang Lehner[1]

[1] Technische Universität Dresden, Database Technology Group, Germany
martin.hahmann@tu-dresden.de, wolfgang.lehner@tu-dresden.de
[2] Robotron Datenbank-Software GmbH, Dresden, Germany
robert.ulbricht@robotron.de
[3] Aalborg University, Denmark
bn21@cs.aau.dk

Abstract. With the ongoing expansion of renewable energy supply, developing and comparing precise forecasting methods becomes important. In this paper, an evaluation metric is investigated which allows the integration of multiple accuracy criteria into one consistent performance ranking and returns information about the economic impact of a forecast. The practical applicability of the approach is demonstrated using solar energy time series observed in different real-world scenarios.

Keywords: Renewable energy forecasting, performance evaluation, ranking score, business context

1 Introduction

The capacity of renewable energy sources like solar panels and wind mills is constantly increasing world-wide. Simultaneously, many countries aim at establishing liberalized electricity markets in order to create competition between the former monopolistic organizations. As a consequence, the maintenance of the electric balance between power demand and supply is challenged (1) technically by the fluctuating character of the renewable sources and (2) economically by the need for a seamless integration of all market participants. Accordingly, a lot of research was dedicated in the past to the development of precise time-series forecasting models for renewable energy supply. However, due to the lack of a industry-wide accepted and standardized evaluation protocol for forecast quality, decisions are primarily based on context-unaware statistical error measures. As such domain-neutral evaluation criteria do not consider the varying economic impact of over- and underestimations at a certain moment, this can result in misleading decisions. Furthermore, by the use of more than one error criterion the obtained ranking for the competing methods can be inconsistent.

The purpose of this work is the introduction of a value-based performance evaluation methodology for renewable energy forecasting methods. Firstly, a time-dependent context component is introduced by the use of electricity spot

market prices to determine the economic benefit obtained from the forecast results in a modeled market environment. Secondly, we propose different forms of an evaluation criterion which combines multiple uni-dimensional accuracy measures in order to solve possible ranking inconsistencies. Finally, by bringing them together we create an integrated context-aware and multi-dimensional approach with the abilities of reflecting the impact of a decision and flexible adaptation to the underlying scenario. The paper contains 5 sections, with the first being this introduction. In Section 2 we discuss state-of-the-art forecasting performance measures used in the energy domain. Subsequently, we provide a basic description of the relevant business environment defined by the core electricity market rules in Section 3 before we introduce our novel approach of context-dependent forecast benefit determination in Section 4. After that, in Section 5 the practical applicability is demonstrated on three real-world use cases and finally, we conclude and outline future research on this topic in Section 6.

2 Accuracy evaluation in energy forecasting

Despite of a wider range of available forecast evaluation criteria like a model's robustness or technical performance, reducing quality determination to the accuracy dimension of a forecast is a common practice in the forecasting community. However, also the selection of appropriate statistical metrics to measure the forecast accuracy is a topic frequently addressed in literature (compare e.g. [2], [3] or [4]). For the renewable energy domain, the foundations of a standardized performance evaluation protocol were defined by Madsen et al. [6] more than a decade ago. As a minimum set of measures, they propose the use of normalized *Mean Bias Error* (MBE), *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE) and the usage of improvement factors for accuracy comparison between concurring methods and against naïve predictors.

Error Term	**Definition**
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{t=1}^{n}\lvert y_t - y'_t\rvert$
Mean Bias Error	$MBE = \frac{1}{n}\sum_{t=1}^{n}(y_t - y'_t)$
Mean Square Error	$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - y'_t)^2$
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - y'_t)^2}$
Mean Absolute Percentage Error	$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left\lvert\frac{y_t - y'_t}{y_t}\right\rvert$
Symmentric Mean Absolute Percentage Error	$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{\lvert y_t - y'_t\rvert}{\lvert y'_t\rvert + \lvert y_t\rvert}$

Table 1. Common statistical error criteria used for measuring forecast accuracy.

When deciding on a specific error criterion, its characteristics have to be taken into account: The simple MAE indicates the magnitude of the average error,

but focuses on the mean which leads to an underrating of high, but infrequent errors. This is corrected by the *Mean Square Error* (MSE), as squaring the error before the mean is calculated puts a higher penalty on large errors. The same information is provided by the RMSE with the exception that the square root brings the result's unit back to the original value. In contrast, the MBE describes the direction of the error bias. It's value is related to the magnitude of the value under investigation. According to the definition in Table 1, a negative MBE occurs when predictions are higher than observations, indicating a systematic over-prediction by the model. As none of these criteria provide information on the relative size of the error, forecasts of different time series can not be directly compared. This is addressed by the *Mean Absolute Percentage Error* (MAPE), which is one of the most popular measures and returns the MAE in percentage terms. However, one of the known shortcomings of MAPE is that it is biased as it will systematically select a method whose forecasts are too low. The reason is that for under-predictions the MAPE cannot exceed 100%, but there is no upper limit for over-predictions. Furthermore, it is not defined for observed zero values. To deal with such limitations the *Symmetrical Mean Absolute Percentage Error* (SMAPE) was proposed [7]. The SMAPE has a lower and an upper bound and can handle zero observations as long as predictions are not zero. Surprisingly, our former research has shown that the SMAPE is rarely found in renewable energy forecasting literature [10].

3 Energy market models

In this section we give a brief introduction to common energy market rules and describe the possible interactions for trading and balancing. We provide details for the three exemplary markets whose most relevant market characteristics are compared in Table 2.

Since the beginning of the process in the 1980s [9], many of the industrialized countries have liberalized their energy markets, thus breaking up with the traditional market roles. Customers can now freely choose their favorite energy supplier and electricity is traded between the market participants in the newly created markets. Similar to other commodities, transactions with a short delivery time are done at the short term spot market and trades that have to be fulfilled further in the future in the long term future market. While a spot market trade is meant to satisfy an urgent need, the motivation of future contracts often is to protect a trading party against the economic risks of unexpected and drastic price movements. In the short-term electricity markets, the purchased power is paid either for one trading day before it is delivered (day-ahead) or, depending on country-specific market rules, for up to x minutes before the delivery at a certain hour (intraday). Prices are fixed through auctions or through continuous trading, although literature [1] suggests that they do not vary that much between day-ahead and intraday contracts.

Once intraday trading is closed, any further deviation in the portfolio is balanced by trading in the regulating power market. The regulating power market is activated shortly before the time of the actual delivery and when the market is anticipated to have any imbalance in supply or demand. The regulating power could be activated for any duration of time. Regulating power can be either up or down as a consequence of the following situations: If the supply is less than the demand, the supplier's associated balancing responsible party (BRP) has to buy up-regulating power - at up-regulating power price - in order to maintain the energy balance in the market. The required amount of up-regulating power is fulfilled by other energy suppliers or by decreasing the demand by an amount equivalent to the difference. On the other hand, if the supply is greater than the demand, the BRP has to sell down-regulating power - at down-regulating power price - to maintain the energy balance in the market. The down-regulating power is sold in the reserve energy market, or the demand is increased by an amount equivalent to the difference. Furthermore, negative wholesale prices can be permitted or penalties can be applied to those who cause such deviations.

	Australia	**Denmark**	**Germany**
Intraday trading	No	Continuous, closure 60min before t_0	Continuous, closure 5min before t_0
Pricing restrictions	$-1,000 \leq P_{SPOT} \leq 330\ AU\$/MWh$	$-500 \leq P_{SPOT} \leq 3,000\ €/MWh$	$-500 \leq P_{SPOT} \leq 3,000\ €/MWh$
Deviation penalties	Cost of regulation power	Cost of regulation power	None for P_{FIT}; Cost of regulation power for P_T option
Regulation prices	$P_{R\_UP} = P_{R\_DW}$	$P_{R\_UP} \neq P_{R\_DW}$	$P_{R\_UP} = P_{R\_DW}$

Table 2. Comparison of selected characteristics for the electricity spot markets of Australia, Denmark and Germany.

(Western) Australia. The Wholesale Electricity Market (WEM) for the South West Interconnected System of Western Australia operates independently from the Australian National Energy Market. Most of the energy trading in the WEM is done directly via bilateral contracts, so in the day-ahead spot market (STEM) only the positions not already covered by such contracts are traded. Positions can be traded until 9:50 AM of each trading day and prices are settled through auctions. After the STEM trades are settled, all deviations from contract positions are exposed to the regulation price. The regulation price is determined according to the minimum and maximum STEM prices for the trading period. Balancing costs are funded by the customers based on their monthly demand [8].

Denmark. The Danish energy market is an integral part of the Nordic energy market, and trading takes place through Nord Pool. The spot market closes at 12 AM, where the market participants submit their bid for power that will be

delivered or purchased on each trading interval of the following trading day. Each trading interval represents an hourly period. Once Nord Pool calculates and informs the market prices to the participants, the trade is settled. Thereafter, any deviation in the commitment to the spot market is handled by bilateral trading or participating in the intraday market which opens at 2 PM and closes 60 min before delivery start. Up- and down regulation power prices are distinct and the cost of regulation is assigned to those participants who are responsible for the imbalance.

Germany. Trading in the German energy market takes place at the European Power Exchange (EEX), where bids for the spot market have to be submitted until 12 PM. Intraday trading starts at 3 PM for the following day, closing 5 min before delivery. Renewable generators can choose between fixed feed-in tariffs P_{FIT} or directly selling their energy to the market and receiving a premium tariff P_T on top of the market price. In case of the latter they are charged for imbalances, otherwise the consumer has to pay the balancing services. In the regulation energy market, only one price is determined for both up- and down-regulating power.

The chosen examples show that for every market design, there are different conditions that can affect the way forecasts are created and evaluated. This also applies to the specific requirements in the forecasting process and to the motivation of the market participants for providing accurate results.

4 Value-based forecasting performance

In this section, first we describe our approach of value-based forecasting performance measurement before we define a metric used for selecting forecasting methods based on ranking scores.

4.1 Forecast Benefit Determination

Something all the error criteria discussed in Section 2 have in common is that none of them is context-dependent. To provide information about the domain-specific economic impact of forecast accuracy when deciding for a specific method, a tailor-made criterion has to be used which allows for modeling the relevant business environment as shown by [5]. In the case of renewable energy suppliers, the economic benefit of a forecast is determined by applying the corresponding electricity market-rules and -prices to the numerical results. This returns a scalar product of two time series which is time-dependent, so whenever there are differences between spot- and regulation-prices the over- and underestimations will be fined differently thus leading to a higher diffusion of the original forecast

accuracy. Subsequently we propose two criteria to measure the benefit for a forecast.

Forecast Value. The *Forecast Value* $FCV(F_n)$ aims at giving information about the absolute monetary return from the day-ahead spot market for a chosen forecasting model F_n (compare Equation 1): The predicted amount of energy y' is always sold at the corresponding electricity spot market price P_{SPOT}. When the actual delivered amount of energy y is higher than anticipated, the surplus energy is bought at 100% by the TSO at the down-regulation price P_{R_DW}. Accordingly, for underproduction the TSO will sell the missing amount to the energy supplier at the up-regulation price P_{R_UP}. In both cases, the expected trading profit can be further increased when premium tariffs P_T are paid on top; or reduced by the deviation penalty D_P.

$$FCV(F_n) = \begin{cases} y' * P_{SPOT} + (y - y') * (P_{R\_DW} + P_T) - D_P & \text{if } y > y' \\ y' * P_{SPOT} + (y' - y) * (P_{R\_UP} + P_T) - D_P & \text{others} \end{cases} \quad (1)$$

Energy producers participating in the spot market will be interested in maximizing their benefit in terms of FCV. This means that depending on their bidding strategy, the most accurate forecast will not necessarily result in the highest FCV. For example when $P_{SPOT} < P_{R\_DW}$, intentional underestimating the scheduled output and selling the surplus energy at the regulation market is more attractive as long as the penalty costs caused by the deviation are lower than the net trading result. However, the FCV can be negative if that bidding strategy fails. An exception are producers that receive fixed feed-in tariffs P_{FIT} and do not have to pay deviation costs, they do not rely on forecast quality because the benefit model for their FCV simplifies to:

$$FCV(F_n) = y * P_{FIT} \quad (2)$$

Forecast Loss. Similar to the FCV introduced above, the *Forecast Loss* $FCL(F_n)$ includes market information but determines the monetary loss for a forecasting model compared to a perfectly fitted result. The FCL is defined as the scalar product of the absolute energy deviation and the difference between spot- and regulation energy prices according to Equation 3:

$$FCL(F_n) = \begin{cases} |(y - y')| * |(P_{SPOT} + P_T - P_{R\_UP})| + D_P & \text{if } y > y' \\ |(y - y')| * |(P_{SPOT} + P_T - P_{R\_DW})| + D_P & \text{others} \end{cases} \quad (3)$$

Hence, the optimal FCL is 0 which means that there is either no error in the forecast or no price difference between spot- and regulation market at the moment of energy delivery. Unlike the FCV, the FCL is insensitive to the bidding strategy of the market participant, although the significance of the time component for forecast accuracy is reflected here as well.

4.2 Multi-dimensional Ranking Scores

Using more than one evaluation criterion when comparing the accuracy of competing forecasting methods is common practice in the energy forecasting domain [10] and in such cases a distinct ranking can be obtained for each criterion. This leads to several inconsistent rankings for different criteria and finally leaves the decision about the optimal method to choose to the user. A multi-dimensional ranking score (e.g. [11]) provides a unique ranking for multiple criteria. In this subsection we introduce the different versions of ranking scores used in this paper.

Absolute Ranking Score. Let n_F be the number of evaluated forecasting methods F and m_E be the number of statistical error measures $E_i(F_n)$ with $1 \leq i \leq m_E$ calculated for each forecast output. Now, for each $E_i(F_n)$ the forecasting methods are ranked starting with $S_i(F_n) = 1$ for the lowest error value $min(E_i)$ to $S_i(F_n) = n_F$ for the highest value $max(E_i)$. The score $S_i(F_n)$ is the rank of F_n for its error measure E_i. The *Absolute Ranking Score* $RS(F_n)$ can then be described as the sum of $S_i(F_n)$ for all respective E_i as shown in equation 4:

$$RS(F_n) = \sum_{i=1}^{m_E} S_i(F_n) \quad \text{with} \quad 1 \leq S_i(F_n) \leq n_F \tag{4}$$

For example, in a setting using 5 different forecasting methods and 3 error measures, the best score that can be obtained is 3 (=first position for each error category), while the lowest score would be 15.

Normalized Ranking Score. As the RS only considers the absolute rank of a forecast in the result list, its scale is quite coarse. The RS provides no information about the magnitude of the distance between one position and the next, so two methods having very close error values would have the same score as two with a much wider spread. The *Normalized Ranking Score* $NRS(F_n)$ corrects this shortcoming. When determining $S_i(F_n)$, the error values of each category are normalized so that $min(E_i) = 0$ and $max(E_i) = 1$. This means that for the $NRS(F_n)$, the optimal value is 0 (lowest value for all E_i) while the worst result is n_F. This way any discrimination is eliminated and furthermore, the probability of having equal ranking scores for methods that simply alternate their absolute result positions is reduced.

$$NRS(F_n) = \sum_{i=1}^{m_E} S_i(F_n) \quad \text{with} \quad 0 \leq S_i(F_n) \leq 1 \tag{5}$$

Weighted Ranking Scores. When choosing number and type of error measures to be used for a ranking, there might be a need to over- or underweight a specific E_m in the final score. This requirement leads to the *Weighted Ranking Score* $WRS(F_n)$ which applies the weighting factor λ_i with $\sum$ to the absolute score $S_i(F_n)$ for each E_i as described in Equation 6. Alternatively, λ_i can be applied to the normalized score as well and is denoted as *Weighted Normalized*

Ranking Score $WNRS(F_n)$.

$$WRS(F_n) = \sum_{i=1}^{m_E} \lambda_i S_i(F_n) \quad \text{with} \quad \sum_{i=1}^{m_E} \lambda_i = 1 \tag{6}$$

Using static or variable weights allows for better adaption of the evaluation metric to specific characteristics of the underlying scenario, thus increasing its' overall flexibility. For example, continuously emphasizing the MBE criteria would lead to better scores for forecasts that do not have a high systematic error, although they might have strong absolute deviations in both directions during the whole evaluation period.

The usage of weighting factors introduces the problem of how to derive the optimal $\lambda_{i,t+1}$ for a given use case in advance. To determine $\lambda_{i,t}$ we use the current rankings for the individual error criteria $R_{i,t}(E_i)$ as the input vectors and one of the corresponding forecast benefit rankings $R_t(B)$ with $B \in \{FCL, FCV\}$ as the target values for the optimization function. For example, if a method F_n had the highest FCV in t, all $\lambda_{i,t}$ are set to values so that the weighted score ranking $R_t(E)$ with $E \in \{WRS, WNRS\}$ shall also return the first position for F_n. However, this will not always be solvable by the optimizer so the quality of the obtained results has to be verified in order to avoid misleading results for λ_i. This is done by calculating the accuracy A_j obtained from the weighted ranking according to Equation 7. The level j of A determines how many positions are relevant for the ranking accuracy, so e.g. setting $j = 5$ means that only the first 5 methods of the ranking are considered while all lower positions will be ignored.

$$a(E,B) = \begin{cases} 1 & \text{if } R(E) = R(B) \\ 0 & \text{others} \end{cases} \quad \text{and} \quad A_j = \frac{1}{j} \sum_{n=1}^{j} a_n \tag{7}$$

For static weights, setting the weighting factors once manually based upon expert knowledge might be suitable. However, as soon as the forecast environment changes in the long term (e.g. by a higher seasonal spread of electricity prices), regular adjustments are necessary to better reflect the new situation. Therefore, we use diurnal persistence thus assuming that the environment of the actual trading interval t is similar or equal to the day before so that for short terms we anticipate $\lambda_{i,t} = \lambda_{i,t-1}$. This decision depends on if $A_{j,t-1}$ is considered as satisfying, e.g. setting a threshold of $A_5 \geq 0.80$ signifies that at least 4 out of the top 5 methods in the list have to be ranked with correct positions, otherwise $\lambda_{i,t-2}$ is used and so on.

5 Evaluation

In this section we evaluate the forecast performance criteria introduced in the Sections 4.1 and 4.2. We describe the characteristics of the used data sets and the methodology of our experiments before we present and discuss the obtained results.

5.1 Methodology

For each of the markets presented in Section 3 we use a data set containing time series with the aggregated measured output from different local solar energy installations. In the Australian scenario, a single roof-top panel installed in Perth is used. For Denmark, the data is taken from a ground-based commercial solar farm close to Aalborg and for the third use case, we have an aggregated measurement of all mixed-type solar panels belonging to a local distribution network covering an area of 46km^2 in central Germany. Up to 4 external influences like e.g. global irradiation and air temperature are taken from an on-site or nearby weather station, and the corresponding prices from the day-ahead spot- and the regulation-market. Note that all time series are measured data, so any deviations possibly caused by unreliable weather predictions are excluded from the results. Furthermore, we extract additional numerical features like the Hour-of-the-day, the Day-of-the-year, and the Clear-sky value from the raw data in order to improve the expected forecast quality for those methods using them. All time series are equidistant without missing values and have a maximum resolution of 60min; at least two years of historical data are provided in all scenarios.

Symbol	**Method**
ARIMA	Auto-Regressive Integrated Moving Average.
ARIMAX	ARIMA with external regressors.
ETS	Exponential smoothing state space model.
GBM	Generalized Boosted Regression Model.
HW	Holt-Winters seasonal exponential smoothing.
KNN	Regression model using weighted k-Nearest Neighbors.
MARS	Multivariate Adaptive Regression Splines.
MLP	Multi-Layer Perceptron. A fully connected feedforward network.
MLR	Multiple Linear Regression
NAIVE_CS	Naive Clear-Sky. Values are taken from Clear-Sky feature (maximum model).
NAIVE_DP	Naive Diurnal Persistence. Values correspond to last day's observation.
NAIVE_ZE	Naive Zero. All values are zero (minimum model).
NNET	Neural Network with a single-hidden-layer.
RF	Random Forest. Regression based on a forest of trees using random inputs.
SVR	Regression model using Support Vector Machines.

Table 3. Forecasting models used for evaluation

Our objective is to measure the forecast benefit for different forecasting methods and compare the outcome to the ranking scores. Therefore, our experiments are organized as follows: For each use case, we apply a selection of 12 competing state-of-the-art forecasting methods (compare Table 3) to predict the future electricity output from the solar energy installations. We include a naïve persistence model *NAIVE_DP* to provide a benchmark as well as a minimum and maximum model (*NAIVE_ZE*, *NAIVE_CS*), to mark the upper- and lower bounds for the expected values. Initially, the first year of historical data is taken for training, and the second year for evaluating the forecasting models. Then, the forecasts are computed on a daily basis using a sliding window over the training data

according to the submission rules for the individual day-ahead spot markets. For example, if a forecast with hourly resolution has to be submitted at 12 AM before the next trading day, the available training history ends with the 11 AM observation and the forecast horizon is 36 hours ahead, of which the first values that still belong to the actual trading day will be discarded. Trading is simulated as day-ahead only, intra-day corrections are not considered. The errors are computed on the forecast output for each interval. First we use the statistical error measures *MAE*, *MBE*, *MSE*, *RMSE*, *MAPE* and *SMAPE* (compare Table 1) as individual criteria and then a combination of all of them to determine the ranking scores *RS*, *NRS*, *WRS*, and *WNRS*. To measure the obtained forecast benefit, FCV and FCL are calculated with the market prices P_{SPOT}, P_{R_UP} and P_{R_DW}.

5.2 Result discussion

The results listed in Table 4 show the monetary advantage when permanently choosing the optimal forecasting model. The lower bound is always marked by the Clear-Sky model *NAIVE_CS*, which means that constantly over-estimating the output would lead to the lowest benefit. On the other hand, for Australia and Germany the highest benefit is obtained when extremely under-estimating using *NAIVE_ZE*, so selling energy at the regulation market brings higher revenues than on the spot market. However, this strategy would not have worked for the Danish market. Although here the spread between best and worst result is the highest with 175%, no naïve model is found among the first ranks so the market rules clearly favor accurate forecast. In contrast, the top-ranked methods in terms of *FCL* are *GBM*, *RF*, *MARS*, *MLP* and *NNET*, which all are sophisticated forecasting methods that make use of external information. In a fluctuating environment, they are more likely to produce forecasts of higher accuracy than uni-variate or naïve methods.

	Australia		**Denmark**		**Germany**	
#	Method	Result	Method	Result	Method	Result
1	NAIVE_ZE	656.27 $	GBM	51,050.61 €	NAIVE_ZE	105,459.87 €
2	ETS	-5.9%	MARS	-2.7%	GBM	-4.2%
3	GBM	-6.9%	MLP	-3.3%	SVR	-4.4%
4	RF	-7.3%	RF	-3.4%	RF	-4.5%
5	MARS	-7.5%	NNET	-4.3%	MARS	-5.3%
...	...	...	...	...	...	...
15	NAIVE_CS	-64.7%	NAIVE_CS	-175.4%	NAIVE_CS	-110.8%
1	GBM	5.00 $	GBM	1,053.29 €	RF	20,467.61 €
2	RF	+35.3%	RF	+2.6%	GBM	+0.8%
3	NNET	+39.4%	MARS	+12.8%	NNET	+4.5%
4	MLP	+42.4%	NNET	+13.5%	MARS	+5.5%
5	MARS	+49.3%	MLP	+33.4%	MLP	+9.2%
...	...	...	...	...	...	...
15	NAIVE_ZE	+3491.4%	NAIVE_CS	+908.5%	NAIVE_ZE	+512.2%

Table 4. Accumulated FCV in the upper- and FCL in the lower part. The first row of each block shows the absolute value, subsequently the percental deterioration is listed.

Figure 1 compares the outcome for the error criteria in terms of FCV on the left, and FCL on the right side. The method selection decision is reconsidered

in a daily interval, so when observation values are available at the end of each trading day, the optimal method from the last period is used again to predict the forthcoming day. It can be seen that for the *FCL*, the ranking scores (red) outperform most of the standard error criteria (light blue). Furthermore, the varying impact of the standard errors can be observed, e.g. basing on the *SMAPE* leads to a higher (avg. +30.6%) benefit for our examples than the *MAPE*. Using the *WNRS* with flexible weights works fine for Australia, for Denmark and Germany *NRS* and *RS* would be preferable. In contrast, for the *FCV* using the standard errors is more convenient than ranking scores with the exception of Australia. Average differences between the different options are much smaller in all scenarios than for the *FCL*.

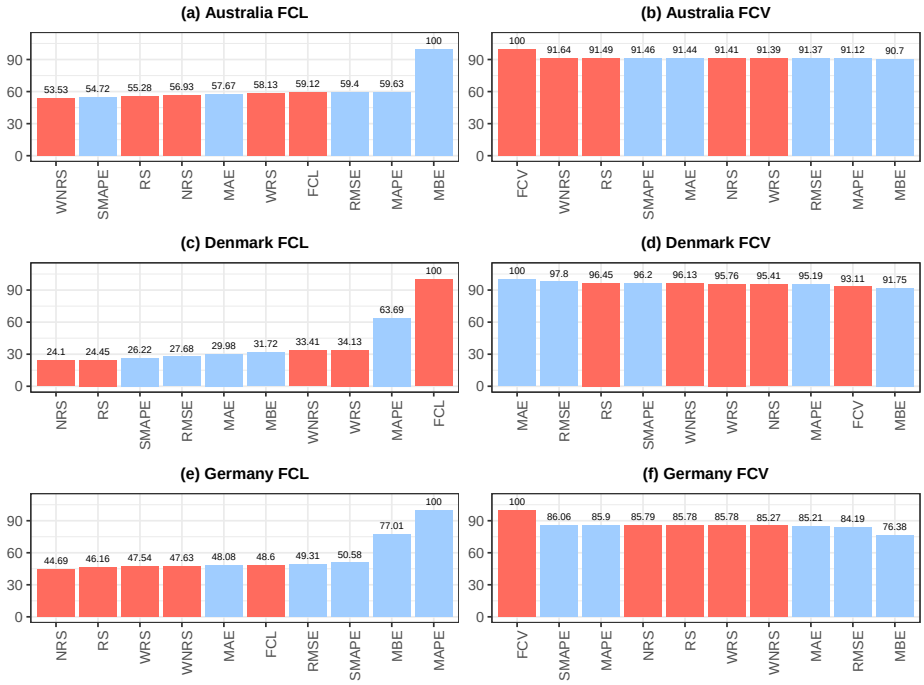

Fig. 1. Impact of selected error measure on total FCL in the left - and FCV in the right column. Standard error measures are displayed in light blue, while the red bars display the specific error criteria introduced in Section 4. Results are normalized and organized in the way that the left bar in each diagram represents the best result obtained.

6 Conclusions

Our findings show that *FCL* and *FCV* are justified context-aware output measures for forecast performance evaluation as both of them illustrate the economic benefit obtained from a specific method. However, they should not be used for

the same purpose. Basing method selection on multi-dimensional ranking scores leads to better forecasting results in terms of a minimized *FCL* than for most of the uni-dimensional criteria but the use of weights does not always outperform unweighted scores. Otherwise, the maximization of the *FCV* is preferably obtained by using the very same *FCV* for markets with one regulation energy price where over- and underestimations are equally fined. In contrast, for varying regulation prices *MAE* and *RMSE* give better results. In fact, the definition of an appropriate evaluation metric strongly depends on the underlying scenario's business context information. Our future work will address refinements of the ranking accuracy measurement and the method selection strategy.

Acknowledgment

The work presented in this paper was funded by the European Regional Development Fund (EFRE) and the Free State of Saxony under the grant agreement number 100269304 and co-financed by Robotron Datenbank-Software GmbH.

References

1. N. Aparicio, I. MacGill, J. Rivier Abbad, and H. Beltran. Comparison of Wind Energy Support Policy and Electricity Market Design in Europe, the United States, and Australia. *IEEE Transactions on Sustainable Energy*, 3(4):809–818, oct 2012.
2. J. S. Armstrong. Evaluating Forecasting Methods. In J. S. Armstrong, editor, *Principles of Forecasting*, volume 30 of *International Series in Operations Research & Management Science*, pages 443–472. Springer US, 2001.
3. Z. Chen and Y. Yang. Assessing forecast accuracy measures. Technical report, Iowa State University, Department of Statistics & Statistical Laboratory, 2004.
4. R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
5. J. Luoma, P. Mathiesen, and J. Kleissl. Forecast value considering energy pricing in California. *Applied Energy*, 125:230–237, jul 2014.
6. H. Madsen, G. Kariniotakis, H. Nielsen, T. Nielsen, and P. Pinson. A Protocol for Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models. Anemos project, European Commission, 2004.
7. S. Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, 1993.
8. I. M. Operator. Wholesale Electricity Market Design Summary. Technical report, Independent Market Operator, 2012.
9. R. Raineri. Chile: Where it all started. In F. P. Sioshansi and W. Pfaffenberger, editors, *Electricity Market Reform: An International Perspective*, pages 77–108. Elsevier, 2006.
10. R. Ulbricht, A. Thoß, H. Donker, G. Gräfe, and W. Lehner. Dealing with uncertainty: An empirical study on the relevance of renewable energy forecasting methods. In *Lecture Notes in Computer Science*, volume 10097 LNAI, pages 54–66. Springer International Publishing, 2017.
11. B. Xu and J. Ouenniche. A multidimensional framework for performance evaluation of forecasting models: context-dependent DEA. *Applied Financial Economics*, 21(24):1873–1890, dec 2011.