A Review of Deterministic Error Scores and Normalization Techniques for Power Forecasting Algorithms

André Gensler, Bernhard Sick
Intelligent Embedded Systems Group
University of Kassel, Germany
Email: {gensler | bsick}@uni-kassel.de

Stephan Vogt
Fraunhofer Institute for Wind Energy and
Energy Systems (IWES), Kassel, Germany
Email: stephan.vogt@iwes.fraunhofer.de

***Abstract*—The evaluation of the performance of forecasting algorithms in the area of power forecasting of regenerative power plants is the basis for model comparison. There are a multitude of different forms of evaluation scores, which, however, do not seem to be universally applied. In this article, we want to broaden the understanding for the function and relationship of different error scores in the area of deterministic error scores. A categorization by normalization technique is introduced, which simplifies the process of choosing the appropriate error score for an application. A number of popular error scores are investigated in a case study which details the development of error scores given different forms of error distributions. Furthermore, the behavior of different error scores on a real-world wind farm data set is analyzed. A correlation analysis between the evaluated scores gives insights on how these scores relate to each other. Properties and notes on the applicability of the presented scores are detailed in a discussion. Finally, an outlook on future work in the area of probabilistic error scores is given.**

NOMENCLATURE

t	Time point of evaluation or *forecasting origin*.
k	Look-ahead time or *forecasting time-step*.
$k_{\max}$	Maximum look-ahead time or *forecasting horizon*.
N	Number of data items used for the evaluation.
$\mathbf{x}_{t+k\|t}$	Model input for time $t+k$ made at time origin t.
y_t	True (measured) power value at time t.
y_{inst}	Installed nominal power capacity.
$\hat{y}_{t+k\|t}$	Forecast for time $t+k$ made at time origin t.
$e_{t+k\|t}$	Forecasting error $e_{t+k\|t} = \hat{y}_{t+k\|t} - y_{t+k}$.

I. INTRODUCTION AND RELATED WORK

Forecasting the power generation of regenerative power plants has been a major area of research due to the increasing portion of regenerative forms of energy in the power grid. These forms of energy have volatile generation characteristics, meaning that the power generation can not be controlled (unlike in conventional power plants). Instead, they depend on the atmospheric conditions, in particular wind speed (and related parameters) for wind turbines, and solar radiation for solar collectors. The typical power forecasting process takes place in two steps:

- A *numeric weather prediction* (NWP) for the desired area is created. This process typically is performed by a weather model provider.
- The NWP (and possibly further data, such as the current power generation) is used to generate a power forecast for a desired time span using a *forecasting algorithm*.

An introduction into the area of forecasting wind power generation is, e.g., given in [1], [2], [3]. Forecasting algorithms typically happened to be physical models, which transform the NWP into a power generation time series using a turbine, or solar collector power curve (e.g., in [4]). However, in the last decade, a multitude of other types of models, e.g., models based on machine learning [5], [6], statistical models [7], or ensemble methods [8], [9], were developed. As the portion of renewable energies in the power grid is steadily increasing, the absolute error may get very large in a worst-case. One of the major goals of research in power forecasting therefore is to develop more sophisticated (i.e., better performing) algorithms.

In order to compare the performance of forecasting algorithms, there has to be a clear definition of how the procedure of quality assessment is performed. There are a multitude of error scores which are utilized, each of which is of interest to a certain participant in the industry, and whose names partially are the same while they are calculated differently. A number of surveys on the forecasting of power generation also include error scores, however, this analysis aims at giving a focused and more complete overview on the topic of forecasting error scores for power forecasting.

Some articles describe and summarize the general assessment of forecasting errors, e.g., [10], [11], [12]. Though these measures are useful and are applied in part in the area of power forecasting, this specific area has certain characteristics which make other forecasting error measures more relevant. The area of the power forecasting has been covered in some surveys, e.g., in [2], [3]. Some other surveys also include sections on forecasting error scores [1], [8], [13], however, they are partially inconsistent with each other and only mention a selection of error scores.

Furthermore, in addition to deterministic errors, the uncer-

TABLE I
OVERVIEW OF BASIC DETERMINISTIC ERROR MEASURES

Error Measure Name	Formula			Purpose
Bias	Bias_k	$=$	$\frac{1}{N_k}\sum_{n=1}^{N_k} e_n$	Shows if an algorithm overestimates or underestimates a forecast (on average).
Mean Absolute Error	MAE_k	$=$	$\frac{1}{N_k}\sum_{n=1}^{N_k} \lvert e_n \rvert$	Linear absolute error measure. Proportional weighting of errors.
Mean Squared Error	MSE_k	$=$	$\frac{1}{N_k}\sum_{n=1}^{N_k} e_n^2$	Quadratic error. Smaller weighting of small errors, larger weighting of large errors.
Root Mean Squared Error	RMSE_k	$=$	$\sqrt{\frac{1}{N_k}\sum_{n=1}^{N_k} e_n^2}$	Square root of MSE has the original physical unit of the forecast.

tainty assessment of a forecast is an increasingly important aspect in power forecasting. It can, but not necessarily has to be expressed as a probability [14]. Having an uncertainty estimate of a power forecast, an actor in the industry has the possibility to plan according to the amount of uncertainty, e.g., taking preventive actions such as increasing the reserve capacity. Reviews on probabilistic error scores for wind power forecasting are described in [15], [16].

The main contributions of this article is a structured overview of existing error measures in the area of deterministic error scores. In two case studies, the characteristics of the presented error measures are analyzed in detail. From the insights of the case studies, advantages and limits in the application of each error score is discussed.

The remainder of this article is structured as follows: Section II summarizes deterministic error scores and categorizes them by their way of normalization. Section III analyzes the presented error scores in a case study to show their effects, which are then discussed in Section IV. Our key findings are summarized in Section V. An outlook on our future work and a short introduction into the area of uncertainty assessment techniques is given in Section VI.

II. DETERMINISTIC FORECASTING ERROR SCORES

A forecast is conducted at a time t, the forecasting origin. Depending on the desired application, a forecast is performed for a number of forecasting time-steps

$$k = k_{\min}, k_{\min}+1, \ldots, k_{\max}. \tag{1}$$

The forecasting time-step borders $k_{\min}$ and $k_{\max}$ can be chosen arbitrarily, for typical applications, such as the day-ahead forecast, the forecasting time-steps are chosen to $k = 24h, \ldots, 48h$. For an intra-day forecast, on the other hand, typical borders are $k = 1h, \ldots, 24h$.

The forecasting process is typically performed using a forecasting model to transform a weather forecast into a power forecast, which in a typical form can be

$$\hat{y}_{t+k|t} = f(\mathbf{x}_{t+k|t}|\boldsymbol{\Theta}), \tag{2}$$

where $\mathbf{x}_{t+k|t}$ are the parameters of a numeric weather prediction at time $t+k$ with forecast origin t, and f is the forecasting model function with model parameters $\boldsymbol{\Theta}$. In case the forecasting algorithm is an ensemble algorithm consisting of J predictions, the deterministic forecast can be computed by

$$\hat{y}_{t+k|t} = \sum_{j=1}^{J} w_j f_j(\mathbf{x}^{(j)}_{t+k|t}|\Theta_{(j)}) \tag{3}$$

with the sum of all weights being $\sum_{j=1}^{J} w_j = 1$. The value of $\mathbf{x}^{(j)}_{t+k|t}$ and the forecasting model parameters $\boldsymbol{\Theta}$ may vary for some forms of ensembles. The *forecasting error* can be calculated after creating the forecast using

$$e_{t+k|t} = \hat{y}_{t+k|t} - y_{t+k}, \tag{4}$$

where y_{t+k} is the power measurement at the corresponding time $t+k$. From this simple form of forecasting error, error measures can be derived.

A. Basic Error Measures

For quality assessment, a number of single deterministic forecasting errors are aggregated into an overall score. There exist a number of scores, which can be seen from Table I. Though there are more sophisticated error scores, most of those scores are based on one of these basic scores.

Each score can either be computed for each forecasting time-step k separately, or as a summarized overall score over all forecasting time-steps. The formula remains the same in this case, though, naturally, all relevant points for the evaluation have to be included then.

The Bias score is just an averaging of all single error values

$$\text{Bias}_k = \frac{1}{N_k}\sum_{n=1}^{N_k} e_n. \tag{5}$$

In itself, this measure has the property of balancing out positive and negative errors. Therefore, it only shows whether an algorithm overestimates or underestimates a forecast on average. The bias itself is not a measure of the forecasting quality of an algorithm, though a low bias is desirable and related to a low error.

The mean absolute error (*MAE*) is computed using

$$\text{MAE}_k = \frac{1}{N_k}\sum_{n=1}^{N_k} |e_n|. \tag{6}$$

The MAE score sums up the absolute error of each forecast. It, therefore, factors in the error distribution in a linear fashion.

TABLE II
OVERVIEW OF COMMONLY USED NORMALIZATION TECHNIQUES

#	Normalization Technique	Formula	Purpose
1	Nominal Capacity	y_{inst}	Scale-free comparison, comparability independent of nominal capacity.
2	Current Power Generation	y_t	Examination of relative error. Errors in low generation scenarios have higher impact.
3	Deviation from Average	$\lvert y_t - \bar{y}\rvert$	Lower weighting of situations at the extreme ends of the generation spectrum.
4	Dynamic characteristics	$\lvert y_t - y_{t-1}\rvert$	Lower weighting of situations with high variability in the power generation.

If the overall minimum difference in error values is to be determined, the MAE is the appropriate score.

The mean squared error (*MSE*) is calculated using

$$\text{MSE}_k = \frac{1}{N_k}\sum_{n=1}^{N_k} e_n^2. \tag{7}$$

Unlike the MAE score, this score factors in the errors quadratically. Thus, high errors are penalized more, while low errors have lower influence on the overall score. If a forecasting algorithm has to avoid extreme errors, the MSE score is the more appropriate error measure. However, the MSE score is a squared score, the value has little relationship with the actual differences. Therefore, this score is mostly used for optimization purposes during forecasting model training. The MSE is optimal during least-squares optimization when assuming a normally distributed error which overlays the deterministic portion of the signal [17].

The root mean squared error (*RMSE*) is computed using

$$\text{RMSE}_k = \sqrt{\frac{1}{N_k}\sum_{n=1}^{N_k} e_n^2}. \tag{8}$$

The RMSE has the same qualitative meaning as the MSE score. However, as the square-root of the MSE value is computed, the value is represented in the original physical unit, making it easier to relate to a forecast value. The reason the MSE is used nevertheless is that it is faster to compute (no computation of the square-root).

B. Normalization Techniques

In the area of power forecasting, there exist a multitude of types of error normalization, each of which has a certain purpose. An overview of the various normalization techniques is given in Table II.

1) The simplest way of normalization is by dividing the forecast value by the *nominal capacity* of the powerplant (i.e., by y_{inst}). The error consequently is computed using
$$\frac{e_{t+k|t}}{y_{\text{inst}}}. \tag{9}$$
This normalization is a constant division factor for each power plant, making it easily understandable. Using this form of normalization, a scale-free comparison of the forecasting quality for different power plants is possible. The overall installed capacity of each power plant is no longer relevant.
2) Another way of normalizing is by dividing the error through the *current power generation* of the power plant y_t, i.e., by calculating
$$\frac{e_{t+k|t}}{y_t}. \tag{10}$$
This form of normalization realizes a relative error in the sense of a percentage error. This type of normalization naturally weighs a certain absolute error in a low power generation scenario higher than in a high power generation scenario (as the percent-wise error is larger).
3) The error can be normalized by factoring in the deviation of a current power generation y_t from the average power generation $\bar{y}$ in the evaluated time span, i.e.,
$$\frac{e_{t+k|t}}{|y_t - \bar{y}|}. \tag{11}$$
This form of normalization penalizes errors near the average power generation, while errors at the extreme ends of the power spectrum have less influence on the overall error score.
4) The error can be normalized with respect to the dynamic characteristics of the current power generation $\Delta y_t = |y_t - y_{t-1}|$, the normalized error consequently is computed using
$$\frac{e_{t+k|t}}{\Delta y_t}. \tag{12}$$
In general, a forecasting problem is more difficult when the dynamics of the weather situation (and thus of the power generation time series) are high. This form of normalization aims to penalize errors in situations with low dynamic variability higher, while situations with high variability are weighted lower. This way of normalization lowers the impact of difficult weather situations.

C. Derived Error Scores

There are a number of additional combined error scores, which are a combination of one of the primary error scores (see Section II-A) and a normalization technique (see Section II-B). The effect of these derived measures consequently is a combination of the primary score and the normalization technique. The derived scores are categorized with respect to their basic score and the normalization technique in Table III. The calculation of the particular scores is again shown in Table IV. Some authors use the same measure name for a score with a different normalization technique (see Table III), therefore, the calculation formula of the precise score should always be given when reporting error scores. Some of the mentioned scores even include multiple normalization techniques.

TABLE III
ERROR MEASURES DEPENDING ON THEIR BASIC ERROR MEASURE AND RESPECTIVE NORMALIZATION TECHNIQUE. THE COMPUTATION OF THE SCORES IS DESCRIBED IN TABLE IV.

#	Measure	Normalization Technique			
		y_{inst}	y_t	$\lvert y_t - \bar{y}\rvert$	$\lvert y_t - y_{t-1}\rvert$
1	Bias	NBias[13]	-	-	-
2	MAE	NMAE[13]	MRE[8], MAPE[8], [12]	-	RAE[12], MASE[11], [8]
3	(R)MSE	NMSE[13], NRMSE[13]	-	NMSE[8], [12], mRSE[12], KL[12]	RSE[12], mRSE[12], U2[12]

D. Deviation Assessment, Correlation and Model Comparison

For deterministic forecasts, it makes sense to not only determine the average error of a forecast, but also the distribution of the errors. The standard measure for deviation assessment is the *standard deviation*. However, as it is already being dealt with errors, the term *Standard Deviation of Errors* (SDE) is introduced in [13], [1], which is nevertheless very similar to the classic standard deviation computation

$$\text{SDE}_k = \sqrt{\frac{\sum_{n=1}^{N_k}(e_n - \bar{e})^2}{N_k - (q+1)}}, \tag{13}$$

where q is the number of estimated parameters using the considered data and $\bar{e}$ is the mean of all error values.

If a more precise assessment of the distribution is desired, the computation of higher moments of the distribution is an option. In particular, the calculation of the *skewness* or *kurtosis* are two reasonable options. Other possibilities to assess the errors are, e.g., using error distribution histograms.

The process of comparing forecasting models is typically done using the *skill score* [13], [1] computed by

$$\text{Imp} = \frac{e_{\text{base}} - e_{\text{eval}}}{e_{\text{base}}}, \tag{14}$$

where e_{eval} is the error score of an evaluated forecasting technique and e_{base} is the error of a baseline technique to compare it to. In many cases, the persistence method or a climatological forecast is used as baseline technique. The result is a factor of improvement Imp, which is positive if the evaluated technique is better than the baseline technique and negative if the baseline technique outperforms the evaluated technique. The skill score therefore often is represented as a percentage value (by multiplying it with 100). It can be applied on any measure, such as MAE, (R)MSE, or even probabilistic scores. It then represents the improvement on the respective score.

Another quality assessment technique is the *coefficient of determination* R^2, which is the squared coefficient of correlation computed by

$$R^2 = \frac{\left(\sum_{n=1}^{N_k}(\hat{y}_n - \bar{\hat{y}})(y_n - \bar{y})\right)^2}{\sum_{n=1}^{N_k}(\hat{y}_n - \bar{\hat{y}})^2 \sum_{n=1}^{N_k}(y_n - \bar{y})^2}, \tag{15}$$

where $\bar{\hat{y}}$ is the average of each forecast value. This measure shows the ability of the model to account for the variance of the data, i.e., it determines the amount of correlation between the evaluated data set and the forecasting model. However, only the amount of linear correlation is assessable, which limits its usefulness. Furthermore, the R^2 score may be high, even though the model may still be incorrect regarding bias or scale. The score is in the range $[0, 1]$. There are a number of scores related to the R^2 score, such as Pearson's r, mutual information, or the Kullback-Leibler divergence, which have a similar meaning. In [13], an additional definition of the R^2 score is given with

$$R^2 = \frac{\text{MSE}_{\text{avg}} - \text{MSE}_{\text{eval}}}{\text{MSE}_{\text{avg}}}, \tag{16}$$

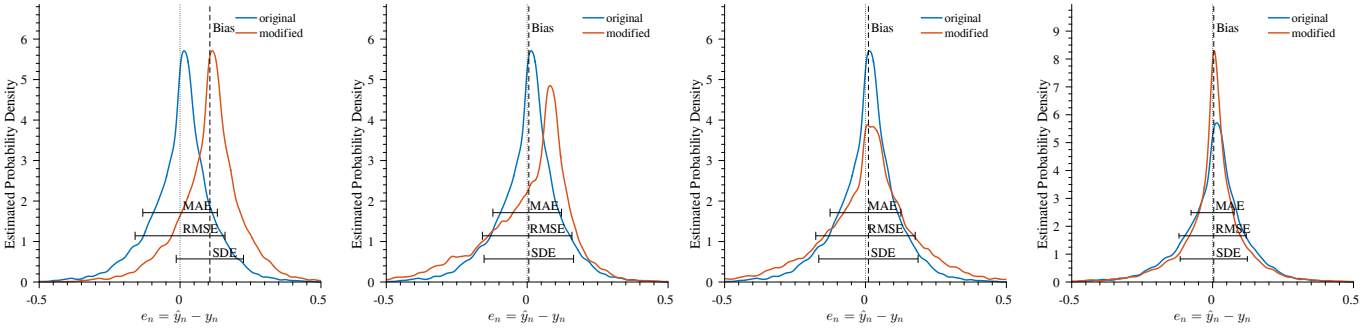
which aims to eliminate some of the disadvantages of the R^2 score. In the above formula, MSE_{avg} is the error of the global mean model (the forecast always is $\bar{y}$). A detailed critique of the R^2 score can be found in [13].

TABLE IV
ADDITIONAL COMBINED FORECASTING ERROR SCORES GROUPED BY NORMALIZATION TECHNIQUE. A CATEGORIZATION OF THESE TECHNIQUES IS SHOWN IN TABLE III.

Measure	Formula
NBias	$\text{NBias}_k = \frac{1}{N_k}\sum_{n=1}^{N_k}\frac{e_n}{y_{\text{inst}}}$
NMAE	$\text{NMAE}_k = \frac{\text{MAE}_k}{y_{\text{inst}}}$
NMSE[13]	$\text{NMSE}_k = \frac{1}{N_k}\sum_{n=1}^{N_k}\frac{e_n^2}{y_{\text{inst}}}$
MRE	$\text{MRE}_k = \frac{1}{N_k}\sum_{n=1}^{N_k}\lvert\frac{e_n}{y_n}\rvert$
MAPE	$\text{MAPE}_k = \text{MRE}_k \times 100\%$
NMSE[8], [12]	$\text{NMSE}_k = \frac{1}{N_k}\sum_{n=1}^{N_k}\frac{e_n^2}{\lvert y_n - \bar{y}\rvert}$
mRSE	$\text{mRSE}_k = \sqrt{\frac{1}{N_k}\sum_{n=1}^{N_k}\frac{e_n^2}{\Delta y_n^2 + \frac{1}{N_k}\sum_{n'=1}^{N_k}(y_{n'} - \bar{y})^2}}$
KL	$\text{KL}_k = \sqrt{\frac{1}{N_k}\sum_{n=1}^{N_k}\frac{e_n^2}{\frac{1}{N_k}\sum_{n'=1}^{N_k}(y_{n'} - \bar{y})^2}}$
RAE	$\text{RAE}_k = \frac{\sum_{n=1}^{N_k}\lvert e_n\rvert}{\sum_{n=1}^{N_k}\lvert\Delta y_n\rvert}$
MASE	$\text{MASE}_k = \frac{1}{\frac{N_k}{N_k-1}}\frac{\sum_{n=1}^{N_k}\lvert e_n\rvert}{\sum_{n=2}^{N_k}\lvert\Delta y_n\rvert}$
RSE / U2	$\text{RSE}_k = \sqrt{\sum_{n=1}^{N_k}\frac{e_n^2}{\Delta y_n^2}}$

III. CASE STUDY

In this section, two case studies aim to show the effects of various error measures. In Section III-A, the behavior of error measures is investigated given a systematical modification of the error distribution the scores are computed on. Section III-B assesses the quality of a number of real-world datasets from a



(1) Model with high bias. (2) Skewed error distribution. (3) Model with higher error spread. (4) Model with different kurtosis.

Fig. 1. Different forms of error distributions. The original distribution is shown by the blue line in each figure. The distribution is then modified to represent a model with a high bias (Fig. 1.1), a skewed error distribution (Fig. 1.2), a higher spread than the original distribution (Fig. 1.3), or a model with a different kurtosis (Fig. 1.4). Basic error scores are listed in each of the error distributions. The behavior of different error scores for each of the presented error distributions is shown in Table V.

wind farm data set using some of the presented scores. Finally, an analysis of the correlation between different error measures is performed.

A. Case Study: Error Distribution Effects

In this case study, we want to investigate the behavior of different error measures given a varying form of error distribution to compute the scores on. The error distribution is taken from a real-world example from a windfarm data set [18]. The error distributions are modified in order to simulate some possible forms of error distributions. The different distributions are shown in Fig. 1. For the sake of better visibility, the distribution is visualized as an estimated density function. We aim to include five forms of modified error distributions in the case study. Besides the original error distribution, a model with a high bias (Fig. 1.1), skewed error distribution (Fig. 1.2), higher error spread (Fig. 1.3), and a different kurtosis (Fig. 1.4) are included. The corresponding error distributions are displayed in Table V.

In the evaluation, we include the most popular error scores, i.e., the basis scores (see Table I), and other popular and frequently used scores, such as SDE, R^2, mRSE, KL, MASE, NMSE [12], and MAPE. All distributions except the biased distribution have no bias, as is shown by the bias score. The values of the RMSE score generally are higher than those of the MAE score due to the high weighting of elements with high distance. This effect can especially be observed if changing the kurtosis: The value of RMSE remains the same, while the MAE value decreases. Both scores represent an increase of model spread proportionally to the scaling factor. The MSE is just the squared distance of the RMSE.

As expected, SDE does not change when changing solely the bias. When having an unbiased model (cases $3-5$), SDE behaves exactly like the RMSE. R^2 drops when having a biased or skewed distribution and is especially sensitive to a higher model spread. However, a change of kurtosis has little impact on the score. The mRSE and KL measures are very close to each other and react very similar to the RMSE error and the inverse of R^2. The temporal normalization Δy_n in mRSE seems to have little impact compared to the second normalization term.

MASE has a different value domain due to the normalization term. However, for the same evaluated data set (as in the present case), the percent-wise change of the MASE corresponds exactly with the MAE error. MAPE is very sensitive to changes of bias, skewness, and kurtosis. It scales linearly when increasing the error spread. The NMSE error has a behavior which is rather hard to interpret. When adding skewness to the error distribution, the NMSE actually decreases (due to high weighting of a number of points close to the average power generation). It is rather unsensitive to changes of bias, skewness, or kurtosis. Interestingly, both MAPE and NMSE do not seem to relate to any of the other scores.

B. Case Study: Quality Assessment for Wind Farms

In a second case study, the behavior of the presented error scores given a real-world dataset of 45 wind farms is analyzed (*EuropeWindFarm* dataset, publicly available at [18]). A normalization is performed for each input dimension.

TABLE V
ERROR SCORES COMPUTED FROM THE ERROR DISTRIBUTIONS OF FIG. 1. THE VALUE OF EACH SCORE IS DENOTED AND THE ABSOLUTE AND RELATIVE CHANGE OF THE RESPECTIVE SCORE FOR EACH OF THE ERROR DISTRIBUTIONS. THE COLORS DENOTE THE SIZE OF THE RESPECTIVE ERROR, WHERE GREEN MEANS LOW ERROR AND YELLOW REPRESENTS A HIGH ERROR.

#	Error Distributions	Bias	%	MAE	%	RMSE	%	MSE	%	SDE	%	R^2	%	mRSE	%	KL	%	MASE	%	NMSE[12]	%	MAPE	%
0	Original distribution	0.006	0	0.085	0	0.120	0	0.014	0	0.119	0	0.740	0	0.481	0	0.510	0	1.834	0	4.34	0	358.6	0
1	Biased distribution	0.106	1721	0.132	56	0.160	33	0.025	78	0.119	0	0.537	-27	0.655	36	0.681	33	2.856	56	2.35	-46	1111.0	210
2	Skewed distribution	0.006	1	0.122	43	0.159	33	0.025	77	0.159	33	0.540	-27	0.644	34	0.678	33	2.629	43	4.38	1	863.8	141
3	More spread (* 1.5)	0.010	80	0.126	48	0.176	48	0.031	118	0.176	47	0.433	-41	0.711	48	0.753	48	2.714	48	9.74	124	536.4	50
4	Different kurtosis	0.004	-29	0.076	-10	0.119	0	0.014	0	0.119	0	0.741	0	0.479	-1	0.509	0	1.650	-10	3.58	-17	285.1	-20

For each wind farm, as forecasting model an extreme learning machine (ELM) is trained using 1500 randomized hidden units and a regularization parameter $\lambda = 10^{-3}$, which are chosen based on expert knowledge. As activation function for the ELM, a rectified linear unit function is chosen. The dataset is split into both, training $(1/3)$ and test dataset $(2/3)$. The results of the experiments are shown in Table VI. The table shows the results of the error scores for each wind farm. The colors denote the relative quality of each measure from low error (green) to high error (red). For an easier inspection of the error scores, the absolute value of the bias is given (A. Bias).

As can be seen from the table, some of the farms are better predictable than others, which can be seen from the values of the various error scores. The basic assumption is as follows: When using the same forecasting algorithm, differences in the quality of the forecast are mostly due to differences coming from the weather situation or power plant location. When comparing forecasting models, a suited score for forecasting model comparison should be able to abstract from these difficulties, i.e., it should have little relative difference. This difference is denoted in the last row of Table VI, which shows the relative percentage between the average value and the standard deviation of each score (shown in the next to last rows). As can be seen, mRSE, KL, and MASE have a smaller relative difference than the established measures such as RMSE, indicating that they might be better suited for model comparison. As can be seen from some of the results, some scores turn out to be problematic in some situations. In particular, this can be observed for the MAPE score (e.g., wf39, wf43) or the NMSE score (e.g., wf4, wf30, wf32), where the scores have massive outliers as results and very big values for the standard deviation.

The correlation of the error measures is further investigated in Fig. 2. The figure shows the absolute correlation of the Pearson correlation coefficient with interval $[0, 1]$. The correlation matrix is computed using the 45 wind farms for each of the 11 investigated error measures. This heat map shows the amount of correlation between the error scores. As can be seen from the figure, the elements along the diagonal line (lower left to upper right) have perfect correlation, as the correlation of each error measure with itself is perfect. The bias has little correlation with the other measures. This is due to the fact that all the investigated trained models have a very small bias, in case of a model with a high bias the correlation to other error measures, such as MAE, does exist. This can, e.g., be observed in the first case study (see Table V). MAE, (R)MSE, and SDE form a central block of high correlation in the figure with correlation higher than 0.9. These measures have no form of normalization which may distort the results, therefore the scores are very understandable. Scores of this category are suited to select a model and give insights on the absolute imprecision of a forecast.

The second category of highly correlated scores are represented by R^2, mRSE, and the KL score. These measures are better suited to compare errors for multiple forecasting

TABLE VI
ERROR SCORES FOR THE EUROPEWINDFARM DATASET. THE COLORS INDICATE THE ERROR SCORE VALUES FROM LOW (GREEN) TO HIGH (RED). THE LAST ROWS DENOTE THE AVERAGE VALUE OF EACH SCORE, THEIR STANDARD DEVIATIONS, AND THE PERCENT-WISE DIFFERENCE.

Data	A.Bias	MAE	RMSE	MSE	SDE	R^2	mRSE	KL	MASE	NMSE	MAPE
wf1	0.011	0.082	0.115	0.013	0.114	0.719	0.497	0.530	1.866	0.328	340.5
wf2	0.032	0.138	0.181	0.033	0.179	0.566	0.609	0.659	1.828	0.647	705.3
wf3	0.000	0.062	0.100	0.010	0.100	0.760	0.441	0.490	1.649	0.439	1033.3
wf4	0.017	0.084	0.116	0.013	0.115	0.755	0.468	0.495	1.806	3.232	425.3
wf5	0.006	0.129	0.207	0.043	0.206	0.306	0.800	0.833	2.662	0.997	451.6
wf6	0.062	0.181	0.286	0.082	0.279	0.197	0.841	1.094	1.576	1.332	312.2
wf7	0.006	0.132	0.182	0.033	0.182	0.685	0.531	0.561	2.007	0.395	157.6
wf8	0.002	0.113	0.158	0.025	0.158	0.702	0.513	0.546	1.948	0.252	160.4
wf9	0.006	0.040	0.071	0.005	0.071	0.561	0.550	0.663	1.677	0.133	292.5
wf10	0.009	0.103	0.138	0.019	0.137	0.719	0.499	0.530	1.948	0.654	421.5
wf11	0.039	0.129	0.189	0.036	0.185	0.639	0.566	0.601	2.147	0.357	385.7
wf12	0.041	0.113	0.165	0.027	0.160	0.651	0.548	0.590	1.859	0.500	1178.0
wf13	0.017	0.063	0.093	0.009	0.091	0.691	0.502	0.556	1.742	0.311	465.1
wf14	0.061	0.103	0.151	0.023	0.138	0.496	0.650	0.710	2.069	0.450	554.5
wf15	0.017	0.079	0.115	0.013	0.114	0.704	0.499	0.544	1.639	0.291	1044.9
wf16	0.006	0.075	0.108	0.012	0.107	0.721	0.481	0.528	1.622	0.233	928.3
wf17	0.010	0.097	0.143	0.020	0.143	0.643	0.552	0.598	1.862	0.650	631.3
wf18	0.001	0.066	0.105	0.011	0.105	0.707	0.498	0.541	1.711	0.290	261.3
wf19	0.002	0.085	0.126	0.016	0.126	0.759	0.458	0.491	1.714	0.243	178.5
wf20	0.019	0.135	0.192	0.037	0.192	0.609	0.587	0.626	2.191	0.423	453.8
wf21	0.008	0.121	0.168	0.028	0.168	0.615	0.568	0.621	1.805	0.350	3670.4
wf22	0.006	0.061	0.090	0.008	0.090	0.695	0.493	0.552	1.668	0.222	1561.8
wf23	0.005	0.099	0.139	0.019	0.138	0.650	0.555	0.592	2.119	0.482	455.9
wf24	0.012	0.077	0.113	0.013	0.112	0.623	0.566	0.614	1.914	0.207	165.2
wf25	0.005	0.087	0.123	0.015	0.123	0.630	0.559	0.608	1.933	0.309	710.8
wf26	0.004	0.103	0.156	0.024	0.156	0.577	0.595	0.650	1.898	0.365	3528.1
wf27	0.003	0.087	0.134	0.018	0.134	0.621	0.539	0.615	1.572	0.353	1282.3
wf28	0.034	0.108	0.170	0.029	0.167	0.636	0.570	0.603	2.441	0.303	5497.2
wf29	0.015	0.063	0.093	0.009	0.092	0.729	0.473	0.520	1.762	0.303	279.1
wf30	0.047	0.131	0.193	0.037	0.188	0.599	0.598	0.633	2.177	3.650	320.0
wf31	0.045	0.168	0.227	0.051	0.222	0.482	0.689	0.719	3.069	0.580	5151.7
wf32	0.008	0.099	0.146	0.021	0.146	0.695	0.506	0.552	1.788	3.635	1496.4
wf33	0.006	0.076	0.113	0.013	0.112	0.650	0.531	0.592	1.820	0.324	827.8
wf34	0.009	0.117	0.160	0.026	0.160	0.691	0.517	0.555	1.799	0.395	355.8
wf35	0.006	0.094	0.133	0.018	0.133	0.804	0.423	0.443	2.109	0.457	1564.3
wf36	0.017	0.096	0.146	0.021	0.145	0.646	0.531	0.595	1.937	0.412	1173.8
wf37	0.008	0.088	0.125	0.016	0.125	0.718	0.484	0.531	1.646	1.043	2591.4
wf38	0.001	0.070	0.114	0.013	0.114	0.717	0.471	0.532	1.747	0.262	439.3
wf39	0.007	0.129	0.173	0.030	0.173	0.728	0.497	0.522	2.223	0.574	54769
wf40	0.034	0.088	0.140	0.020	0.136	0.658	0.535	0.585	1.898	0.483	403.4
wf41	0.005	0.065	0.114	0.013	0.114	0.504	0.603	0.705	1.892	0.237	342.7
wf42	0.023	0.158	0.206	0.043	0.205	0.526	0.634	0.689	1.819	0.765	126.2
wf43	0.010	0.083	0.118	0.014	0.117	0.835	0.384	0.406	1.691	0.324	20047
wf44	0.013	0.116	0.158	0.025	0.158	0.684	0.528	0.562	1.931	0.779	432.9
wf45	0.020	0.162	0.218	0.047	0.217	0.698	0.534	0.549	3.113	0.362	1179.9
Avg.	0.016	0.101	0.147	0.023	0.145	0.636	0.544	0.594	1.940	0.652	2639.0
Std.	0.016	0.031	0.042	0.014	0.041	0.156	0.083	0.107	0.331	0.799	8431.2
%	100	30.7	28.6	60.9	28.3	24.5	15.3	19.6	17.1	122.5	319.5

models. For instance, the value of these scores can indicate whether there remains a potential for improvement, or if the best possible outcomes are already achieved. As can also be seen, there exists a moderate correlation to the block formed by MAE, (R)MSE, and SDE.

The MASE score forms an own category of error score. The form of normalization aims to reduce the influences of the respective location and from the observed time period. MASE still has a moderate correlation to the first block of error scores which can be expected, as it measures the absolute distance.

MAPE and NMSE have almost no correlation with any of the other error measures. Both scores employ a form of normalization which can lead to singularities in this application (as situations with very little power generation or a power generation close to the average generation are very likely to exist in the dataset). The error scores can, thus, be influenced dramatically by a very small number of individual errors. The problematic nature of these scores is further supported by the extreme development of the scores in the observed dataset. The fact that these scores have very little correlation with MAE, (R)MSE, and SDE shows that NMSE and MAPE struggle to correctly assess the quality of a forecast correctly in this application. As these scores are also based on some form of error, a certain amount of correlation *should* exist.

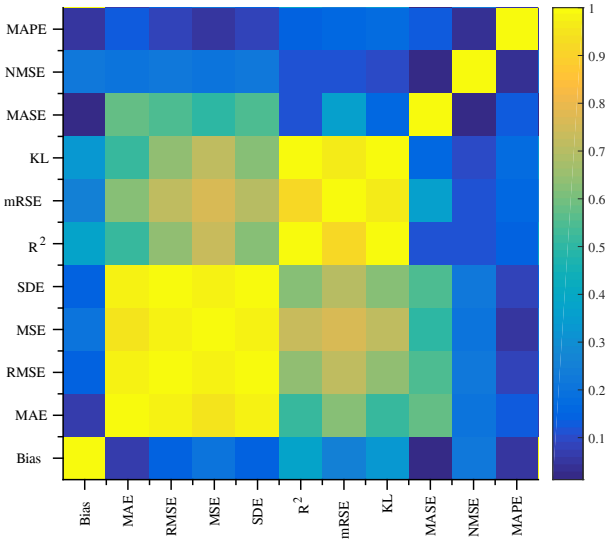
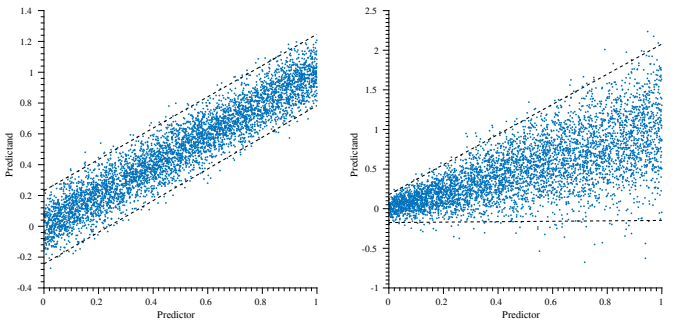


Fig. 2. Absolute value of the Pearson correlation coefficient displayed in a matrix. Naturally, each score has perfect correlation with itself, as is displayed by the yellow diagonal line. A large block of correlated scores is formed by MAE, (R)MSE, and SDE. A second block of high correlation contains R^2, mRSE, and the KL score. MASE has moderate correlation with MAE, (R)MSE, and SDE. NMSE and MAPE have no correlation with any of the other scores.



(1) Homoscedastic uncertainty. (2) Heteroscedastic uncertainty.

Fig. 3. Different forms of uncertainty estimation. In Fig. 3.1, a homoscedastic uncertainty estimation is performed, the uncertainty intervals are equal throughout the predictor input space. Fig. 3.2 shows a heteroscedastic uncertainty estimation. This form of uncertainty estimation is able to identify the change of the amount of uncertainty depending on the predictor input space.

IV. Discussion of Deterministic Error Scores

This section discusses the use of deterministic error scores. In Section IV-A, general properties of the employment of error scores are discussed. Section IV-B deals with the handling of different forms of normalization. Section IV-C discusses how to deal with multiple time horizons and deviation assessment.

A. General Error Score Properties

The optimal score heavily depends on the desired application and market participant. Considering the basic scores, the RMSE should always be preferred to the MSE score when presenting results, as the error units are better understandable. However, for model training the MSE score is equivalent (and faster to compute). RMSE and MAE are equally important, the more appropriate score depends on the target application (or the target audience in the industry, respectively). For electricity trading (where a deviation typically has linear costs), the MAE is preferable. The use of cost/reward functions or loss functions makes sense in conjunction with the MAE when the monetary consequences have a direct relationship to the error of a forecast and are the center of interest. This may, e.g., be the case when an error fee has to be paid for each *kWh* of deviation between forecast and observation. The function can be defined asymmetrically for power surplus or power deficit; furthermore it can be designed in a nonlinear fashion. For grid operators and other grid stability oriented market participants, the RMSE score is more appropriate (as the problematic nature of extreme errors is reflected more appropriately).

The use of skill scores makes sense to compare models with each other, however, they can be misleading if compared to very weak models, such as a climatological model. In any case, the precise computation of all used error scores should be given, as some of the score names are defined in different ways.

B. Handling of Normalization

Considering the forms of normalization, again, their application depends on the desired outcome. Normalization by y_{inst} is very unproblematic and can (and should) be conducted in every case for model comparison. While the idea of the normalization with y_n is elegant for model assessment (percentwise error), it has the unwanted property of a singularity when having a power generation near the low end. Thus, scores based on this normalization technique can eventually be dominated by a small number of measurements where the power generation happens to be very low (as could be observed by the MAPE score in Table VI). Beyond unequal weighting, the division through y_n can turn out problematic numerically (as y_n may be 0). Therefore, we think error measures using this form of normalization (such as the popular MAPE score) should define some form of lower bound ρ for the denominator which results in a limit for maximum weighting, such as

$$y_{n,\mathrm{limit}} = \max(y_n, \rho). \tag{17}$$

As an estimate, we think the value of ρ should be chosen in the range of $0.05 \leq \rho \leq 0.2$ assuming y_n is in the range $[0, 1]$.

The idea of the normalization with $|y_n - \bar{y}|$ is to weigh extreme power generation situations lower. When computed on a data set, it aims to statistically penalize data sets who have little variation in the data. While the idea is commendable, it has the same singularity in the normalization (see NMSE [12] in Table VI), thus should be treated again with some form of threshold limit, such as performed by Eq. 17.

We think the normalization by Δt makes sense in filtering out the impact of highly dynamic and thus difficult weather situations and the effects of "step" errors, i.e., errors which occur when misjudging the point in time of a sudden ramp in the power generation. An approach for this type of normalization is also discussed in [19]. Thus, it has a similar goal

than normalization technique $|y_n - \bar{y}|$, however, it seems to be more tailored towards time series (all other normalization techniques basically can be computed for standard regression problems as well). Again, in itself, this form of normalization has the same disadvantages as the two previous normalization techniques.

However, when the normalization terms are summarized, such as in mRSE, KL, RAE, or MASE, the normalization lowers the impact of the *overall* weather situations which occurred in the data set in the evaluated period. In addition, this form of normalization eliminates the impact of singularities and therefore the need of thresholding, such as performed by Eq. 17. Especially when comparing algorithms which were evaluated on different data sets, these scores can help to make the algorithms better comparable. We therefore think scores based on this form of normalization is preferable for this task.

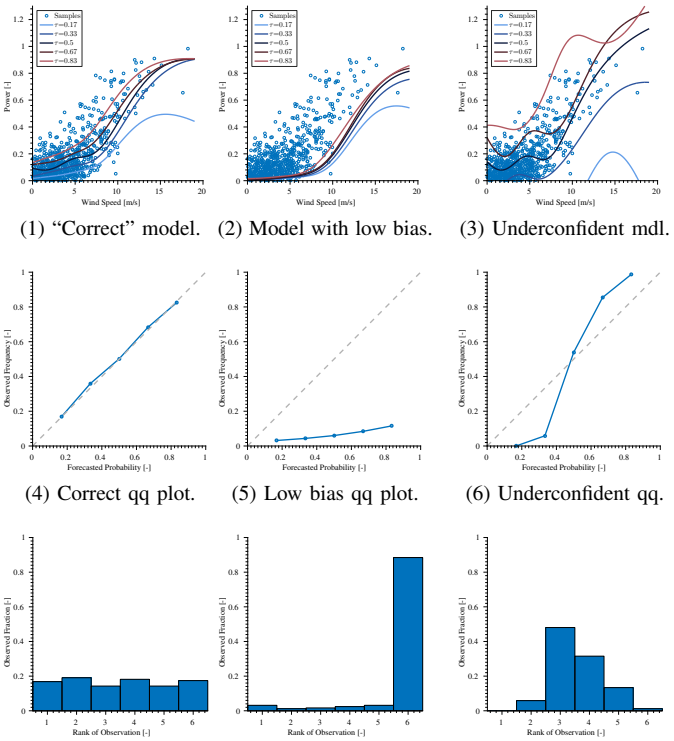
C. Multiple Time Horizons and Deviation Assessment

For some applications, it may make sense to aggregate the errors not for each forecasting time-step, but as a whole (i.e., one score for the whole forecasting time-span). Typically, in the literature this is performed just by computing the score while including the errors from all time horizons. However, as errors typically increase when forecasting time-steps which are further away from the forecasting origin, errors close to the forecasting horizon may dominate the overall result. Especially for (very-)short term algorithms, such as the persistence method, this turns out disadvantageous. Though not practiced on a regular basis, an option worth investigating is the use some form of time horizon smoothing, e.g., in the form

$$\mathrm{RMSE} = \frac{1}{k_{max} - k_{min} + 1} \sum_{k=k_{min}}^{k_{max}} \frac{\mathrm{RMSE}_k}{k}, \qquad (18)$$

when aggregating forecast errors for multiple forecasting time-steps.

Even for deterministic scores we think measures such as the SDE should be determined in the evaluation, as it gives an insight on the anticipated fluctuation of errors. A disadvantage of this form of deviation assessment is the implicit normal distribution assumption of the error, which may not always hold. Furthermore, the error distribution is computed over the complete evaluated data set, assuming an equal error distribution over the data set (i.e., homoscedasticity assumption), which is usually not the case. This form of uncertainty representation is visualized in Fig. 3.1. However, this form of uncertainty representation would fail to correctly assess the uncertainty in case of the uncertainty distribution of Fig. 3.2. Situation-dependent uncertainties can be assessed using *probabilistic uncertainty assessment techniques* which are heteroscedastic, i.e., they are able to assess uncertainty unequally for different areas of the input space. This form of uncertainty representation is shown in Fig. 3.2. While more difficult to assess, heteroscedastic uncertainty usually fits the real uncertainty in a power forecasting context more precisely, as, e.g., the absolute error of a forecasting algorithm is higher in a situation with a lot of wind than in low-wind situations. A small introduction into the area of probabilistic uncertainty assessment techniques is given in Section VI.



(1) "Correct" model. (2) Model with low bias. (3) Underconfident mdl.

(4) Correct qq plot. (5) Low bias qq plot. (6) Underconfident qq.

(7) Correct reliability. (8) Biased rank hist. (9) Too much spread.

Fig. 4. Different visualization techniques for uncertainty assessment. The figure shows three quantile regression models (row 1) with their respective qq plots (row 2) and rank histograms (row 3). The leftmost figures show a model where the assumed distribution fits the observed distribution, the middle column shows a model with low bias, and the right model shows an unbiased, but underconfident model.

V. Conclusion

This article presents some of the most frequently used deterministic error scores and gives insights on how to use them depending on the desired application. In the area of deterministic error scores, we highlight some of the commonly used normalization techniques, which simplifies the categorization of different forms of errors. In two case studies, we investigated the behavior of the most popular error scores depending on the form of the error distributions they were computed on, and their correlation. In a discussion, we gave some insights and best practices on the use of advanced error measures.

VI. Outlook

In the future, we aim to use the presented deterministic scores for more sophisticated ensemble variants which make use of quantile regression techniques. We also intent to include some of the presented error scores as objective function for machine learning models during model training.

Moreover, we plan to expand the analysis of error scores to the assessment of uncertainty and their representations,

which are, e.g., described in [15], [16]. We aim to present different forms of uncertainty representations, such as interval forecasting techniques, quantile regression, or probability density functions. In this context, we want to assess the role and necessity of ensemble techniques which may be a prerequisite for some forms of uncertainty assessment. Therein, we will highlight different methods of ensemble creation. We then plan to investigate forms of assessing the uncertainty both numerically and visually.

The following example gives a short insight in how a visual inspection can work given a discrete uncertainty representation, here in the form of a quantile regression model created by repeated model training of extreme learning machines using a modified loss function for each quantile during model training. Popular forms of visual uncertainty assessment techniques can, for instance, be so-called quantile-quantile (qq) plots, and Talagrand diagrams, often also referred to as rank histograms.

A qq plot shows the relation of assumed quantile positions to the actual observed quantiles in the data. Ideally, the quantiles are located on the diagonal line, which means the estimated quantiles match the observed quantile positions. Three examples of qq plots are shown in the middle row in Fig. 4. A qq plot for a model with correct estimation of the quantiles (Fig. 4.1) is shown in Fig. 4.4. The model of Fig. 4.2 shows a model with a "low" bias, as can be seen by the line which is located under the ideal line in Fig. 4.5. Furthermore, it can be observed that the model is overconfident, as the slope of the qq plot is lower than the ideal slope. Fig. 4.3 shows an underconfident model, it has too much spread. This can be observed in Fig. 4.6 by the s-curve in the qq plot.

The rank histogram is a simplified form of assessment of uncertainty. Assuming the quantile intervals are equally distributed (e.g., $0.2, 0.4, 0.6, 0.8$), the rank histogram is optimal if the same amount of observations is in every bin, such as can be seen from Fig. 4.7. In the same fashion as the qq plot, in Fig. 4.8 a low bias can be seen as there are too many observation in the highest bin. Fig. 4.9 shows a variant with too much model spread, which can be seen by the little number of samples which are in the outer bins. In the future, we aim to investigate the techniques presented above and further uncertainty assessment techniques. We intent to highlight their advantages and disadvantages. We also plan to give hints on which uncertainty assessment techniques are advantageous for which forms of uncertainty representation depending on the form of the used forecasting model or ensemble.

ACKNOWLEDGMENT

This article partly results from the project Big Energy (HA project no. 472/15- 14), which is funded in the framework of Hessen Modell-Projekte, financed with funds of LOEWE Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

REFERENCES

[1] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, no. 1, pp. 1–8, 2012.

[2] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *North American Power Symposium (NAPS)*, 2010, pp. 1–8.

[3] M. Lei, L. Shiyan, J. Chuanwen, L. Hongling, and Z. Yan, "A review on the forecasting of wind speed and generated power," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 4, pp. 915–920, 2009.

[4] U. Focken, M. Lange, and H.-P. Waldl, "Previento - A Wind Power Prediction System with an Innovative Upscaling Algorithm," in *Proceedings of the European Wind Energy Conference (EWEC)*, 2001, pp. 1–4.

[5] R. L. Welch, S. M. Ruffing, and G. K. Venayagamoorthy, "Comparison of Feedforward and Feedback Neural Network Architectures for Short Term Wind Speed Prediction," in *International Joint Conference on Neural Networks*, 2009, pp. 3335–3340.

[6] T. Barbounis and J. Theocharis, "Locally recurrent neural networks for wind speed prediction using spatial correlation," *Information Sciences*, vol. 177, no. 24, pp. 5775–5797, 2007.

[7] A. J. Conejo, M. a. Plazas, R. Espínola, S. Member, and A. B. Molina, "Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models," *IEEE Transactions On Power Systems*, vol. 20, no. 2, pp. 1035–1042, 2005.

[8] Y. Ren, P. Suganthan, and N. Srikanth, "Ensemble methods for wind and solar power forecasting - A state-of-the-art review," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 82–91, 2015.

[9] A. Gensler and B. Sick, "Forecasting Wind Power - An Ensemble Technique With Gradual Coopetitive Weighting Based on Weather Situation," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN16)*, Vancouver, Canada, 2016, pp. 1–9.

[10] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting*, vol. 16, no. 4, pp. 437–450, 2000.

[11] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[12] Z. Chen and Y. Yang, "Assessing forecast accuracy measures," *Preprint Series*, pp. 1–26, 2004.

[13] H. Madsen, P. Pinson, and G. Kariniotakis, "Standardizing the performance evaluation of shortterm wind power prediction models," *Wind Engineering*, vol. 29, no. 6, pp. 475–489, 2005.

[14] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: A survey of theories and practices," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2463–2482, 2013.

[15] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[16] J. Jung and R. P. Broadwater, "Current status and future advances for wind speed and power forecasting," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 762–777, 2014.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY, USA, 2006, vol. 4, no. 4.

[18] A. Gensler, "EuropeWindFarm Data Set," 2016. [Online]. Available: http://ies-research.de/Software

[19] J. Dobschinski, *Vorhersage der Prognosegüte verschieden großer Windpark-Portfolios*. Kassel, Germany: Intelligent Embedded Systems, Kassel University Press, 2016.