Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models

Henrik Madsen[1], Pierre Pinson[2], George Kariniotakis[2], Henrik Aa. Nielsen[1] and Torben S. Nielsen[1]

[1] Technical University of Denmark, Informatics and Mathematical Modelling, 2800 Lyngby, Denmark.

[2] Ecole des Mines de Paris, Centre for Energy and Processes, 06904 Sophia-Antipolis, France. Lead author, E-mail: hm@imm.dtu.dk

ABSTRACT

Short-term wind power prediction is a primary requirement for efficient large-scale integration of wind generation in power systems and electricity markets. The choice of an appropriate prediction model among the numerous available models is not trivial, and has to be based on an objective evaluation of model performance.

This paper proposes a standardized protocol for the evaluation of short-term wind-power prediction systems. A number of reference prediction models are also described, and their use for performance comparison is analysed. The use of the protocol is demonstrated, using results from both on-shore and offshore wind farms. The work was developed in the frame of the Anemos project (EU R&D project) where the protocol has been used to evaluate more than 10 prediction systems.

Keywords: Wind power forecasting, prediction error, performance evaluation, evaluation protocol.

NOMENCLATURE

P_{inst},	Wind farm installed capacity (in kW or MW)
$k = 1, 2, \ldots, k_{\max}$,	Prediction time-step (also called lead time or look-ahead time)
$k_{\max}$,	Maximum prediction horizon
N,	Number of data used for the model evaluation
$P(t)$,	Measured power at time t (in kW or MW), which usually corresponds to the average power over the previous time period
$\hat{P}(t+k\|t)$,	Power forecast for time $t+k$ made at time origin t (in kW or MW)
$e(t+k\|t)$,	Error corresponding to time $t+k$ for the prediction made at time origin t (in kW or MW)
$\varepsilon(t+k\|t)$,	Normalized prediction error (normalized to the installed capacity)
χ_e,	Random error
μ_e,	Systematic error (bias)
BIAS(k),	Systematic error (estimate) for the prediction horizon k.
MAE(k),	Mean Absolute Error
MSE(k),	Mean Squared Error
RMSE(k),	Root Mean Squared Error
STD(k),	Standard Deviation

NBIAS, NMAE, NMSE, NRMSE, NSTD. Normalized error measures. Calculated using the normalized prediction error.

1. INTRODUCTION

Short-term forecasting of windfarm power production, up to 48 hours ahead, is recognized as a major contribution for reliable large-scale wind power integration. Increasing the value of wind generation by improving the performance of the prediction systems is identified as one of the priorities within wind energy research [1]. Especially in a liberalized electricity market, prediction tools enhance the position of wind energy compared to other forms of dispatchable generation. The increasing need for wind power has encouraged various industrial companies and research organisations to produce such forecasting tools.

In the last decade, and particularly at conferences (e.g. Global Windpower 2002, EWEC 2003, etc.), several prediction tools have been presented [2–11]. This has enabled important feedback from end-users showing the need for standardized methodology for evaluating the accuracy of prediction models.

The performance of each prediction system depends on both the modelling approach and the characteristics of the intended application. Nowadays, due to the cost of prediction systems, and to the economic impact that their accuracy may have, there is a clear demand by end-users for a standardized methodology to evaluate model performance.

This paper presents a complete protocol, consisting of a set of criteria appropriate for the evaluation of a wind-power prediction system. This protocol is a result of the work performed in the frame of the Anemos Project, where the performance of more than 10 prediction systems were evaluated on several on-shore and offshore case studies [12]. The Anemos project is a European R&D project on short-term wind power prediction. It aims at developing accurate models for on-shore and offshore wind resource forecasting using statistical as well as physical approaches. As part of the project, an integrated software system, 'Anemos', has been developed to host the various models. This system will be installed by several utilities for on-line operation at on-shore and offshore wind farms for both local and regional wind-power prediction. The project includes 23 partners from 7 countries.

To develop this evaluation protocol, about 150 suitable references were studied in detail for criteria on wind power prediction. Difficulties with some of the currently used error-measures are briefly mentioned here. Recent examples have shown, especially when there is commercial interest, that standard statistical criteria are often not used correctly, so, giving erroneous conclusions about the accuracy of a given model [13].

Furthermore, a set of simple models is introduced for reference predictors. These, include persistence, the global mean, and a new reference model. They provide a basis for comparison with more advanced models. Example results are given on a real case study. Finally, guidelines are produced for the use of the set of criteria proposed in this paper.

2. PROPOSED SET OF ERROR MEASURES

In this section, we introduce the notation that is commonly used in the wind power forecasting community. Then, the reference models are presented and the definitions of the proposed set of error measures are given.

2.1 Notation of this Section

P_{inst} : Wind farm installed capacity (in kW or MW)

$k = 1, 2, \ldots, k_{\max}$: Prediction time-step (also called lead time or look-ahead time)

$k_{\max}$: Maximum prediction horizon

N : Number of data used for the model evaluation

$P(t)$: Measured power at time t (in kW or MW), which usually corresponds to the average power over the previous time period

$\hat{P}(t+k|t)$: Power forecast for time $t+k$ made at time origin t (in kW or MW)
$\bar{P}(t)$: average of all the available observations of wind power up to time t
$e(t+k|t)$: Error corresponding to time $t+k$ for the prediction made at time origin t (in kW or MW)
$\varepsilon(\mathrm{t}+k|\mathrm{t})$: Normalized prediction error (normalized to the installed capacity)

2.2 Reference Models

It is worthwhile developing and implementing an advanced wind power forecasting tool if this improves upon reference models, especially if by simple considerations and not by increase of modelling effort. Probably the most common 'reference model' used for wind power prediction and meteorology is 'Persistence'. This naive predictor states that future wind production remains the same as the last measured value of power, i.e.

$$\hat{P}_p(t+k\,|\,t) = P(t). \tag{1}$$

Despite its apparent simplicity, this model might be hard to beat for the first look-ahead times (say up to 4-6 hours). This is due to the scale of changes in the atmosphere, which are in general slow. A generalization of the Persistence model is to replace the last measured value by the average of the last n measured values

$$\hat{P}_{MA,n}(t+k\,|\,t) = \sum_{i=0}^{n-1} P(t-i)/\mathrm{n}. \tag{2}$$

Such models are sometimes referred to as 'moving average predictors'. Asymptotically (as n goes to infinity), they tend to the global mean

$$\hat{P}_0(t+k\,|\,t) = \bar{P}(t), \tag{3}$$

The Global Mean can also be seen as a reference model, but since it is not dynamic, its performance may be very poor for the first prediction horizons. However, for longer look-ahead times, its accuracy is much better than Persistence. The performance of these two reference models has been analytically studied by Nielsen et al. [14]. Consequently, the authors proposed to merge the two models in order to get the best of their performance over the whole range of prediction horizons. The merging yields a new reference model

$$\hat{P}_{NR}(t+k\,|\,t) = a_k P(t) + (1-a_k)\overline{P(t)}, \tag{4}$$

where $\mathrm{a_k}$ is defined as the correlation coefficient between $P(t)$ and $P(t+k)$.

All the important statistical quantities, namely $\bar{P}(t)$ and a_k (k = 1,2..., $k_{\max}$), must be estimated or fixed using the training set of data, c.f. also the discussion in the following Section.

2.3 Training and Test Data

When setting up a prediction model, the first step is to take decisions on the structure of the model (e.g. how many neurons for a neural network) before estimating the model parameters based on the available data. The next step should provide a measure for the model performance that will characterize its quality. The quality of a model can be objectively assessed only on a test set of data, which must be independent of the dataset previously used for model building and training. The capability of a model to perform well when it predicts new and independent data is defined as 'generalization performance'.

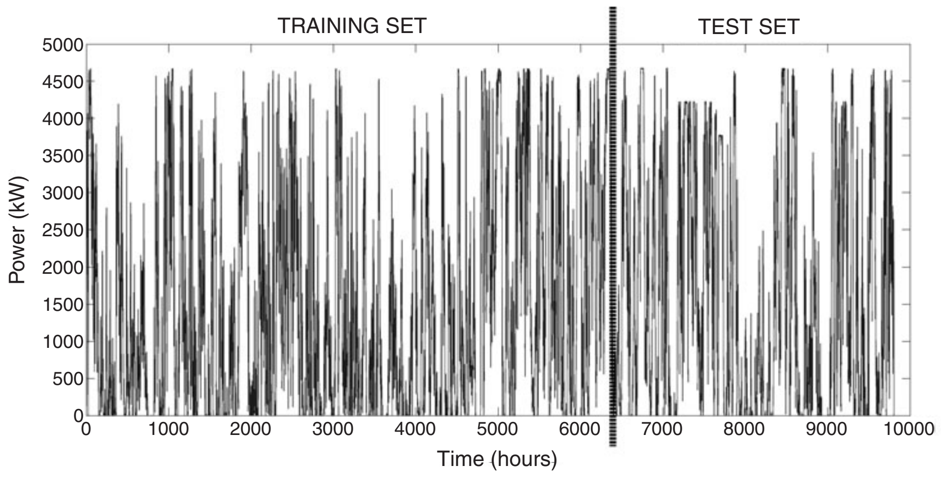

Figure 1: A data set from the offshore wind farm Tunø Knob in Denmark, split into an initial training period and subsequent test period.

It is thus essential to evaluate measures for the errors in prediction, as will be proposed in the next section. The data used for evaluation must not have been used for setting up the prediction model or for tuning its parameters. For this reason, the available data must be split into a 'training period' and a 'test period', as illustrated in Figure 1. Some procedures for model building need a validation set for decisions about the model structure, for instance by cross-validation. Any such validation data are a part of the training set. Error measures resulting from the training set are called 'in-sample' measures; error measures resulting from the test set are called 'out-of-sample' measures.

It is emphasized that training (or estimation) error does not provide a good estimate of the test error, which is the prediction error on new (independent) data. Training error consistently decreases with model complexity, typically dropping to zero if the model complexity is large enough. In practice, however, such a model is expected to perform poorly. This can be concluded from the performance of the model on the test set.

Hence, it should be clear for model developers and for end-users that training data are only dedicated for initially tuning the model, even if very good performance can be reported on this set. Reported error measures must be based on the test period only. Furthermore, it should be ensured that the evaluation made on this set mimics the operational application of the model.

2.4 Definition of Error Measures

2.4.1 Prediction Error

In the field of time series prediction in general, the prediction error is defined as the difference between the measured and the predicted value[1]. Therefore, since we consider separately each forecast horizon, the *prediction error* for the lead time t+ k is defined as

$$e(t+k\,|\,t) = P(t+k) - \hat{P}(t+k\,|\,t)\,. \tag{5}$$

Often it is convenient to introduce the *normalized prediction error*, by the installed capacity P_{inst}, for comparison among errors referring to different wind farms

[1] Intuitively however a positive prediction error refers to over-prediction. The error is then defined as $e_I(t+k\,|\,t) = \hat{P}(t+k\,|\,t) - P(t+k)$.. This can be used if one wants to monitor errors for visualization purposess

$$\varepsilon(t+k\,|\,t) = \frac{1}{P_{inst}} e(t+k/t) = \frac{1}{P_{inst}} (P(t+k) - \hat{P}(t+k/t)). \tag{6}$$

Any prediction error can be decomposed into systematic μ_e error and random error χ_e, viz.

$$e = \mu_e + \chi_e\ , \tag{7}$$

where μ_e is a constant and χ_e is a zero mean random variable.

2.4.2 Definitions of Error Measures

The model bias, which corresponds to the systematic error, is estimated as the average error over the whole evaluation period and is computed for each horizon

$$BIAS(k) = \hat{\mu}_e = \bar{e}_k = \frac{1}{N}\sum_{t=1}^{N} e(t+k\,|\,t). \tag{8}$$

There are two basic criteria for illustrating a predictor's performance: the Mean Absolute Error (abbreviated *MAE*) and the Root Mean Squared Error (abbreviated *RMSE*). The Mean Absolute Error is

$$MAE(k) = \frac{1}{N}\sum_{t=1}^{N} \left| e(t+k\,|\,t) \right|\ . \tag{9}$$

Before introducing the *RMSE* it is useful to introduce the Mean Squared Error (MSE)

$$MSE(k) = \frac{1}{N}\sum_{t=1}^{N} e^2(t+k\,|\,t)\ . \tag{10}$$

The Root Mean Squared Error is then simply

$$RMSE(k) = MSE^{1/2}(k) = \left(\frac{1}{N}\sum_{t=1}^{N} e^2(t+k\,|\,t) \right)^{1/2}\ , \tag{11}$$

Notice that both systematic and random errors contribute to the *MAE* and *RMSE* criteria.

An alternative to the use of the *RMSE* is to consider the Standard Deviation of Errors (*SDE*):

$$SDE(k) = \left(\frac{1}{N-(p+1)}\sum_{t=1}^{N} (e(t+k\,|\,t) - \bar{e}_k\,)^2 \right)^{1/2}\ , \tag{12}$$

where p denotes the number of estimated parameters using the considered data. Hence for the test data $p = 0$. It is noted that similar correction with the number of parameters p should be included in the previous criteria when they refer to the training set of data.

The *SDE* criterion is an estimate for the standard deviation of the error distribution, and then only the random error contributes to the *SDE* criterion.

Statistically the values of *BIAS* and *MAE* are associated with the first moment of the prediction error, and hence these are measures which are directly related to the produced energy of the wind farm. The values of *RMSE* and *SDE* are associated with the second order moment, and hence to the variance of the prediction error. For these latter measures, large prediction errors have the largest effect.

All the error measures introduced above can be calculated using the prediction error $e\,(t+k|t)$ or the normalized prediction error $\varepsilon(t+k|t)$. The purpose of using normalized error measures is to produce results independent of wind farm sizes. The resulting error measures are then referred to as Normalized *BIAS* (*NBIAS*), Normalized *MAE* (abbreviated *NMAE*), and so on.

Here, the proposed normalization is made by the installed capacity, P_{inst}. This contrasts with the possibility to provide the error as a percentage of measured (or predicted) power. This alternative is not obviously feasible since a measured power value may equal zero. Alternatively, if the prediction error is evaluated over a long period, it is then possible to normalize the considered criterion by the average measured power production *P* over the whole period. For the example of the *MAE*, this yields

$$NMAE_m(k) = \frac{MAE(k)}{\overline{P}} = \frac{\sum_{t=1}^{N} |e(t+k \mid t)|}{\sum_{t=1}^{N} P(t)} \quad . \tag{13}$$

This mode of normalization allows better assessment of the monetary consequences of the model errors as a function of the capacity factor of the wind farm (e.g. of inaccurate forecasts of electricity generation within a liberalised electricity market).

Some references use other definitions of error measures. One example is the so-called 'power surplus' for a given period, which is the sum of all positive prediction errors; likewise the 'power deficit' is the sum of all the negative errors over the test period.

2.4.3 Comparison of Models

When evaluating an advanced model, it is important to quantify the benefit of the advanced approach compared to the reference. This gain, denoted as an ``improvement" with respect to the considered reference model, is defined as follows for a given lead time:

$$I_{ref,EC} = 100.\frac{EC_{ref}(k) - EC(k)}{EC_{ref}(k)} \ (\%) \ , \tag{14}$$

where *EC* is the considered Evaluation Criterion, which can be either *MAE*, or, *RMSE* or, even, *SDE*, or the equivalent normalized versions.

Another way to illustrate the skill of advanced forecasting methods is to compute the coefficient of determination R^2 for each look-ahead time:

$$R^2(k) = \frac{MSE_0(k) - MSE(k)}{MSE_0(k)} \tag{15}$$

where $MSE_0(k)$ is the Mean Squared Error for the global mean model (cf. eqn (3)) where the average is estimated upon the available data.

The coefficient of determination represents the ability of the model to explain the variance of the data. The value of R^2 is between 0 for useless predictions and 1 for perfect predictions.

The R^2-value is designed for model selection using the training set, and we suggest avoiding the use of this criterion as a main tool for performance evaluations. If, for instance, the naive predictor is used for large horizons, the resulting R^2-value will be negative! This is because the asymptotic variance of the prediction errors for the naive prediction is twice the variance of the global mean prediction defined by eqn (3) [14]. The R^2-value can be considered for comparing the performance of various models, and/or for various sites, but then it should be remembered that this is out of the scope of its primary use.

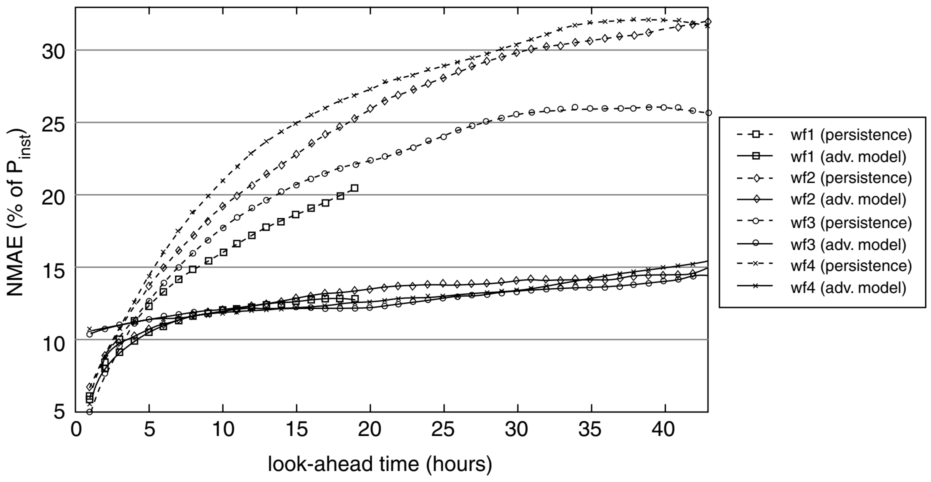

Figure 2: Performance in terms of NMAE of two predictors (Persistence and a state-of-art artificial-intelligence based prediction method) for four different sites.

There exist several alternative definitions of the R^2-value. One frequently used is based on the correlation between the measured and predicted wind power. The problem of this definition is that, even though the predictions might be biased both with respect to level and scale, this definition may lead to $R^2 = 1$. The definition given by eqn (15) does not pose this problem, since both the systematic and random error are embedded in the *MSE* values. Thus, if the R^2-value is reported it is extremely important to describe exactly how it is calculated.

2.5 Factors Influencing the Value of Error Measures

Apart from the capacity of a forecasting method itself, both characteristics of the site and period of time covered by the test set may also significantly influence the apparent performance of a given forecasting system. Figure 2 illustrates that comment. It depicts an evaluation of Persistence and of an advanced approach performances for 4 different sites. The advanced approach is a state-of-the-art artificial-intelligence based forecasting method. Performance is assessed with the *NMAE* criterion. The four windfarms span the various possibilities for site characteristics: complex terrain (wf1), semi-complex terrain (wf2), flat terrain (wf3) and offshore conditions (wf4). From the plot, one can notice that for wf3 and wf4, which are both located in Denmark with similar meteorological conditions, the advanced model performance differs by approximately 20% (i.e. ordinate scale, 2 percent points in 10).

In a recent paper, Kariniotakis et al. [12] compare, in a systematic way, the performance of several prediction models for various case-studies with different characteristics. It is shown how site characteristics, and more precisely terrain complexity, may affect the prediction error, whatever the forecasting method used.

3. EXPLORATORY ANALYSIS

Several other criteria can be used for exploratory analysis. Here, we present some of the methods which are found to be of particular interest in relation to wind power prediction. These tools for exploratory analysis of the prediction errors provide deeper insight into the performance of the methods.

A histogram plot showing the distribution of prediction errors is useful, since it contains more information concerning the error dispersion than a single criterion like the *SDE* or *RMSE*.

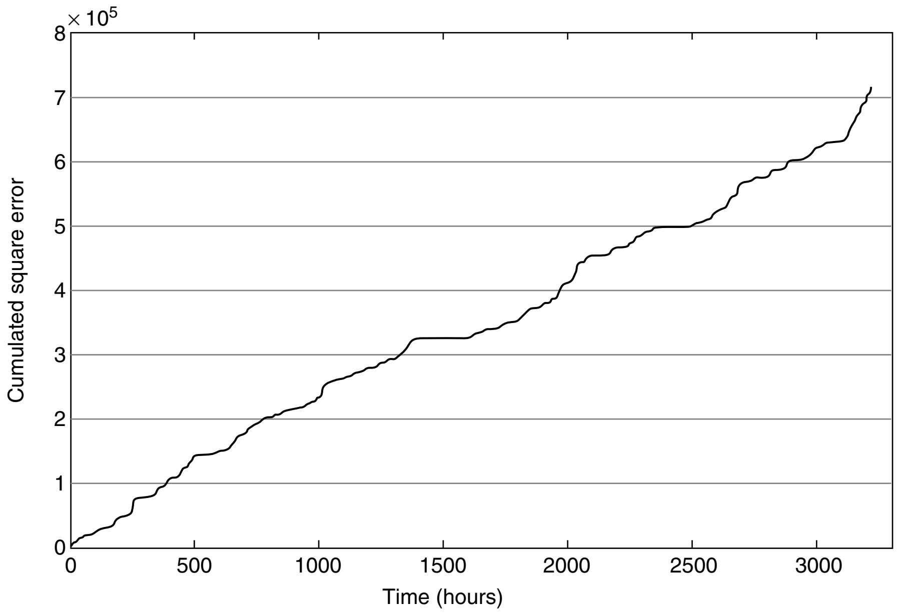

Figure 3: Cumulated squared prediction errors for the wind power production at the Tunø Knob offshore wind farm.

Various conclusions can be derived from both preliminary observations based on visual inspection (such as symmetry, skewness, tails/outliers) and from numerical treatment of error distributions. It should, however, be noticed that the errors are not stationary, and hence the histogram could be plotted as a function of the expected condition, e.g. strong wind speed, summer, westerly wind, etc. This may permit to detect and analyse a model behaviour related to specific conditions. An example of using the histogram will be shown for the case study considered in Section 4.

Another useful tool is a plot of the cumulated squared prediction errors. When using the evaluation criteria (of the previous section) to estimate the general performance of a prediction model over a given period, the cumulated squared errors exhibit the dynamical behaviour of the model performance. For example, the method detects both (i) changes in the Numerical Weather Predictions (NWPs) used as input, and (ii) problems with the model auto-adaptation scheme. Also, if several models are compared with that measure, periods for which certain models perform better than the others can be easily spotted.

For instance, Figure 3 depicts the cumulated measure for 6 hour predictions for the Tunø offshore wind farm. The plot shows a clear change in the increment for the cumulated squared prediction errors for the last two weeks of the considered period; recognising such a change should lead to further investigations.

4. APPLICATION TO A REAL CASE STUDY

As an illustration error measures, the case study of a real multi-MW wind farm located in Ireland is considered. A state-of-the-art statistical prediction model is used to provide hourly predictions for a two-day ahead horizon, using Hirlam NWPs and on-line production data as input. NWPs are provided 4 times per day at the level of the wind farm as interpolated values. The forecasting model is evaluated over a 3-month period corresponding approximately to Winter 2003.

Figure 4 depicts the normalized bias (*NBIAS*) as a function of the look-ahead time, showing values between –0.14% and 0.01% of the wind farm installed capacity. Practically, this means that for this case study, the model does not make any significant systematic error. This is a desired property when using a prediction model. Nowadays, both statistical and physical models enhanced with Model Output Statistics (MOS) are able to provide unbiased forecasts.

Figure 5 illustrates the performance evaluation by the use of both the *NMAE* and the *NRMSE*. The two error measures are computed for the advanced model and for the reference one (Persistence is used here), for every prediction horizon. The *NMAE* can be interpreted: straightforwardly; for instance, the advanced model made an average error representing 13%

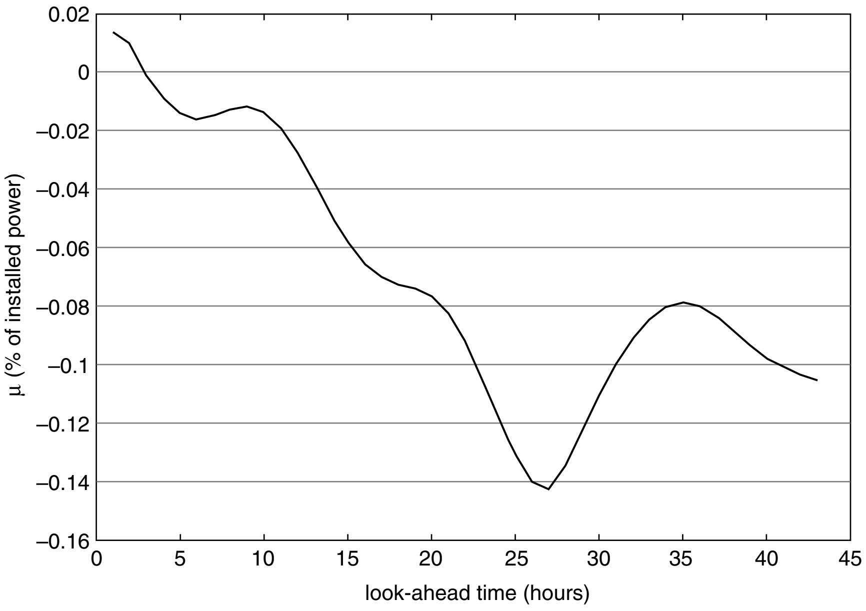

Figure 4: Prediction model bias as a function of the lead time.

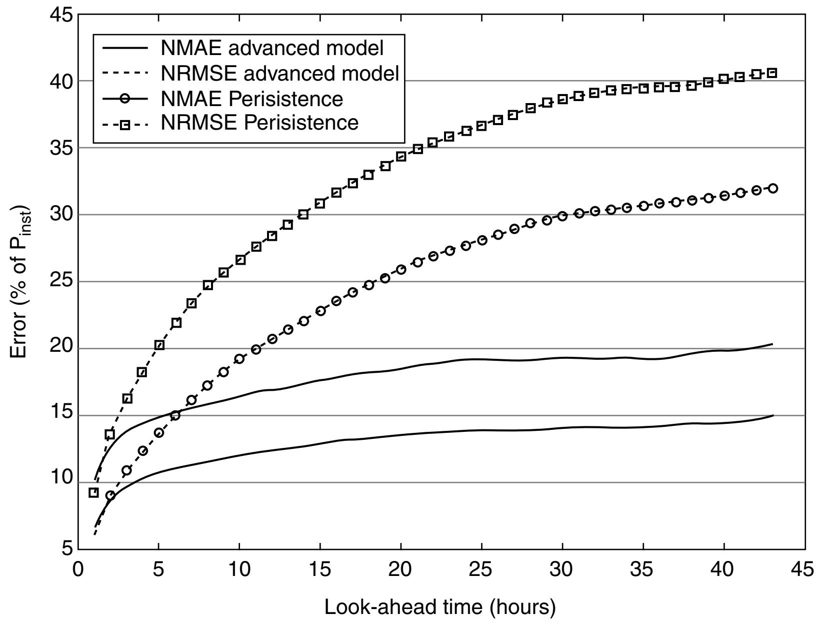

Figure 5: Use of the NMAE and the NRMSE for assessing the performance of the advanced prediction approach, and for comparison with a reference predictor (Persistence is used here).

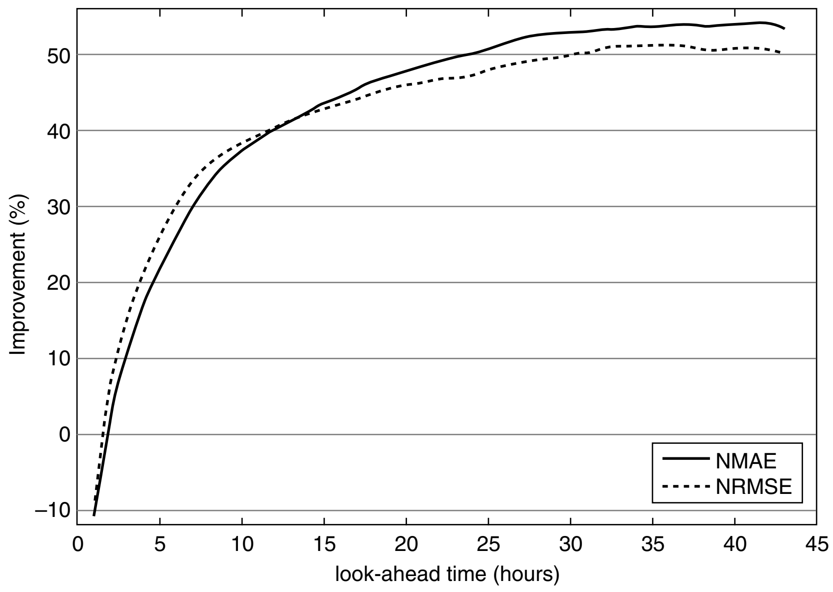

Figure 6: Improvement with respect to Persistence for both the NMAE and the NRMSE criteria.

of the installed capacity for one-day ahead predictions, over the whole evaluation period. Such information is not provided by the *NRMSE* because it considers squared errors and thus gives more weight to large errors. The *NRMSE* measure is more relevant if the aim is to study the impact of these large errors.

The model's benefit is then assessed by calculating the error reduction it achieves against a reference model (Figure 6). Persistence is considered here. An advanced prediction approach should give significant improvement over simple reference models, in order to justify the modelling efforts involved in their design. Here, the improvement for both criteria ranges from –10% for the first look-ahead time, to almost 55% for longer-term predictions. Beating Persistence for the first horizons is not easy, although for longer-term (12–48 hour ahead), large improvements can be achieved. Then, in order to emphasise the difference between various advanced approaches over the whole range of horizons, the new reference model introduced above should be preferred.

Finally, more subtle information can be extracted from error distributions, as shown in Figure 7. They are produced for the 1st and 24th lead times, with bins representing 5% of the installed capacity. A first inspection at the histogram sharpness, skewness and inf/sup bounds, already gives a good idea of the model's performance. Comparing the two histograms of Figure 7, notice that the error distributions are almost perfectly symmetric and centred around 0, and that the error distribution of the one-hour ahead predictions is much sharper than the other. During the evaluation period, the model never made errors greater than 40% of the installed capacity for the first lead time. This is not the case for 24-hour ahead forecasts.

Scott [15] suggested that the optimal range w for histogram bins is related to the range of the data (range(e)) and the number of samples N as follows:

$$w = \frac{range(e)}{\log_2(N)+1} \tag{16}$$

However, since this proposition may lead to large bins, it is recommended to define a bin size representing 5% (like for the case of Figure 7) or 10% of installed capacity. All bins must have the same size in order to avoid misleading interpretations of the error distributions.

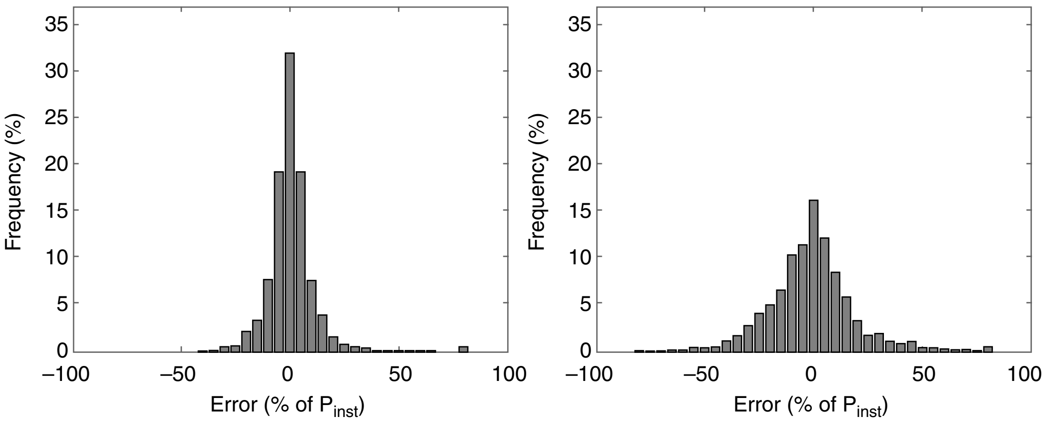

Figure 7: Normalized prediction error distributions for the first look-ahead time (left) and for lead time 24 (right).

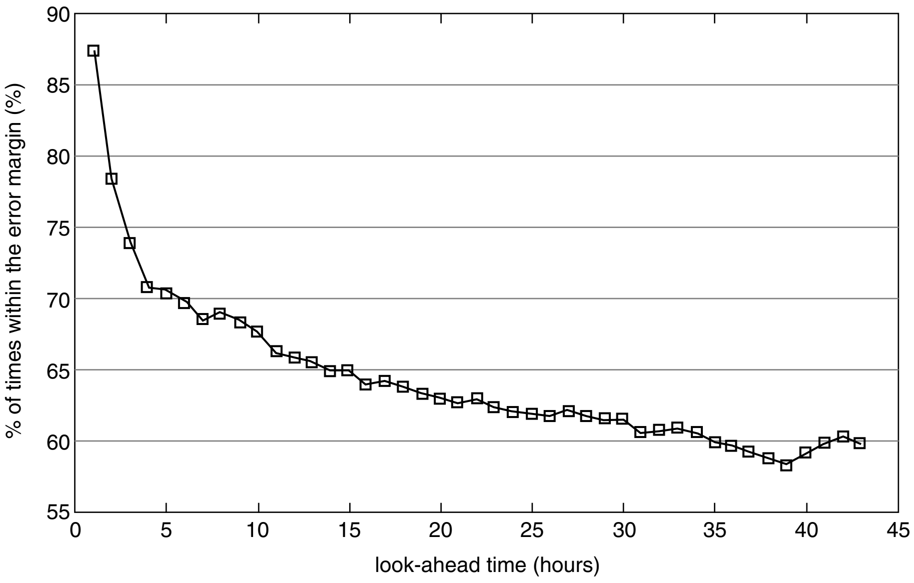

Figure 8: Forecast accuracy of the advanced model in terms of percentage of times the forecasting errors are within a given error margin (+/−12.5% of P_{inst}).

Histograms allow one to quantify the frequency of occurrence of errors below or above a certain level. Examples of some conclusions that can be derived for the considered case study are:

- Robustness
  - 1 step-ahead predictions: errors are less than 7.5% of P_{inst} 68% of the times,
  - 24 step-ahead predictions: errors are less than 7.5% of P_{inst} 24% of the times,
- Large errors
  - 1 step-ahead predictions: errors are more than 17.5% of P_{inst} only 3% of the times,
- etc.

One way to summarize such statistics is to plot the percentage of time that errors are within a given margin. Figure 8 gives an example of that measure as a function of the look-ahead time. Here, an error margin of +/-12.5% of the installed capacity is considered. The use of this measure allows one to estimate the loss in forecast accuracy as the lead time increases. As shown in the figure, the forecast accuracy sharply decreases for the first lead-times and then tends to stabilize. For high horizons, the predictions remain within the error margin with a high probability (higher than 60%).

The combination of all these error measures gives a global view of the ability of a prediction model. Characterization of the errors is not only a primary requirement for end-users to select prediction systems, but also for modellers when developing research towards the improvement of these models.

5. GUIDELINES AND RECOMMENDATIONS

This section contains guidelines and recommendations for providing error measures when evaluating models for short term prediction of wind energy.

5.1 Recommendations

Regarding the performance measures, we have the following recommendations:

- Initially, define clearly the operational framework as discussed in the next section.
- Base performance evaluation on the test set only. The length and period (beginning/end) of the test set should be clearly defined. Moreover, an assessment of the quality of the considered data (i.e. detection of missing or erroneous data) should be performed before starting with the performance evaluation.
- As a minimum set of error measures, the following should be used:
  - NBIAS
  - NMAE
  - NRMSE
- Use the improvement scores for comparison between models.

This is a suggested minimum set of measures. Other measures and tools for exploratory analysis might be used in addition. These measures should be given per time step. Given the variability of the performance of a prediction model, it is useful to provide these measures not only over the whole test set but also for sub-periods (i.e. per month). The values of the measures should be given for both advanced methods and selected simple reference models.

Finally, it should be realized that the most appropriate measure depends on the intended application. Indeed, energy managers and energy traders do not use wind power forecasts in the same way. Hence they often have different views on the cost of prediction errors. This paper supports the analysis of prediction models in terms of power (MW). When this step is rigorously performed following the proposed protocol, one can apply more criteria, based on specific market rules to evaluate the monetary cost of prediction errors. Presenting such criteria is out of the scope of this paper since they usually depend on the particular electricity market.

5.2 Operational Framework

Before presenting any performance measure, it is very important to specify the operational framework of a prediction model. A description of the operational framework includes a specification of:

- Installed capacity. Number and type of wind turbines.
- Horizon of predictions (1, 2, ..., 48, .. hours ahead).

- Use of on-line measurements as input. Specify which data are used (e.g. power production, wind speed, etc.).
- Sampling strategy. Specify whether the measurements are instant readings or the average over some time period, e.g. the last 10 minutes before the time stamp. This should be specified for all observed variables.
- Characteristics of NWP forecasts (frequency of delivery, delay in delivery, horizon, time step, resolution, grid values or interpolated at the position of the farm).
- Frequency of updates of the prediction model. Actually, some models only give forecasts when new NWPs are delivered (i.e. every 6, 12 or 24 hours) while some others operate with a sliding window (typically one hour) since they consider on-line data as input.

6. CONCLUSIONS

Nowadays, there is a great need for standardizing error measures and reference models for assessing the performance of advanced approaches for wind power prediction. Comparison with Persistence does not give a fair measure of the performance of an advanced model, since even the use of the global mean as predictor leads to a 50% reduction in the variance of the error compared to the error obtained with Persistence [14].

This paper introduces guidelines for evaluating wind power predictions, as well as a minimum set of suggested error measures. It is emphasized that a rigorous use of data is required; both training and test sets should be clearly defined and separated.

In the description above, we have focused on prediction models for single wind farms. The modifications needed for considering the wind power predictions for larger areas are minor, given that the relevant measurements are available.

Beside the set of recommended error measures, the researcher should perform further (exploratory) analyses of the prediction errors, e.g. comparisons with other (simple) predictors, histograms, plots of cumulated squared errors, etc. This allows a deeper understanding of the limitations of a given method and indicates possible improvements.

Also, when the performance of a prediction model is evaluated for a given application, it should be specially tailored to the end-user needs. To complement the minimum set of error measures described in this paper, additional criteria representing the monetary cost of the errors may be considered.

The presented set of measures is mostly designed for off-line evaluations. Some of the measures might also be used in on-line situations, e.g. for performance monitoring. The performance measures presented here should be differentiated from methods recently developed for on-line estimation of the uncertainty of wind power predictions [16–18] (prediction intervals for instance). In the later part of the Anemos project, we will elaborate on performance measures which focus on an evaluation of forecast uncertainty estimates. This will be a subject of increasing interest for future research dealing with on-line wind power predictions.

The sequence of prediction errors is obviously correlated, and the so-called autocorrelation of such time-series might be of importance. This holds in particular for wind power generators having auxiliary generation, e.g. from energy storage and able to use the prediction models to schedule the auxiliary plant.. Hence, an operational approach for presenting the autocorrelation of the error sequence is needed; this subject is also dealt within the frame of the Anemos project.

ACKNOWLEDGEMENTS

This work has been partly funded by the European Commission under the 5th Framework Program, in the frame of the Anemos project (contract ENK5-CT-2002-00665). The authors are grateful to the partners involved in the project for providing the necessary data for the realization of the present study.

REFERENCES

1. Thor S.E. and Weis-Taylor P., *Long-term Research and Development Needs for Wind Energy for the Time Frame 2000–2020,* Wind Energy, 2003, 5, 73–77.
2. Landberg L. and Watson S.J., *Short-term Prediction of Local Wind Conditions*, Boundary Layer Meteorology, 1994, 7, 171–195.
3. Bailey B., Brower M.C. and Zack J., Short-term *Wind Forecasting: Development and Application of a Mesoscale Model*: Proceedings of the 1999 European Wind Energy Conference and Exhibition, Nice, France, March 1999, 1062–1065.
4. Nielsen T.S., Madsen H. and Tøfting J., *Experiences with Statistical Models for Wind Power prediction*: Proceedings of the 1999 European Wind Energy Conference and Exhibition, Nice, France, March 1999, 1066–1069.
5. Madsen H., Nielsen T.S., Nielsen H. Aa. and Landberg L., *Short-term Prediction of Wind Farm Electricity Production*: Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering, Barcelona, Spain, September 2000.
6. Focken U., Lange M. and Waldl H.P., Previento – *A Wind Power Prediction System with an Innovative Upscaling Algorithm*: Proceedings of the 2001 European Wind Energy Conference and Exhibition, Copenhagen, Denmark, July 2001, 826–829.
7. Marti I., Nielsen T.S., Madsen H., Navarro J. and Barquero C.G., *Prediction Models in Complex Terrain*: Proceedings of the 2001 European Wind Energy Conference and Exhibition, Copenhagen, Denmark, July 2001, 875–878.
8. Nielsen T.S., Madsen H., Nielsen H.Aa., Landberg L. and Giebel G. Zephyr – *The Prediction Models:* Proceedings of the 2001 European Wind Energy Conference and Exhibition, Copenhagen, Denmark, July 2001, 868–871.
9. Kariniotakis G. and Mayer D., *An Advanced On-line Wind Resource Prediction System for the Optimal Management of Wind Parks*: CD-Proceedings of the 2002 MedPower Conference, Athens, Greece, November 2002.
10. Sánchez I., Usaola J., Ravelo O., Velasco C., Domínguez J., Lobo M., González G.M. and Soto F., SIPREOLICO – *A Wind Power Prediction System Based on a Flexible Combination of Dynamic Models. Application to the Spanish power system*: CD-Proceedings of the 2002 World Wind Energy Conference, Berlin, Germany, June 2002.
11. Giebel G., Kariniotakis G. and Brownsword R., *The State of the Art in Short-term Prediction of Wind Power*, Technical report, Deliverable report of the EU project Anemos, 2003, available: http://anemos.cma.fr.
12. Kariniotakis G. and Marti I., What Performance Can Be Expected by Short-term Wind Power Prediction Models Depending on Site Characteristics ?: CD-Proceedings of the 2004 European Wind Energy Conference and Exhibition, London, United Kingdom, November 2004.

13. Nielsen C.S. and Ravn H.F., Criteria in Short-term Wind Power Prognosis: CD-Proceedings of the 2003 European Wind Energy Conference and Exhibition, Madrid, Spain, June 2003.

14. Nielsen T.S., Joensen A., Madsen H., Landberg L. and Giebel G., *A New Reference Model for Wind Power Forecasting*, Wind Energy, 1998, 1, 29–36.

15. Scott D.W., *Multivariate density estimation: Theory, practice and visualization*, Wiley, New York, 1992.

16. Karniotakis G. and Pinson P., *Uncertainty of Short-term Wind Power Forecasts* – A Methodology for On-line Assessment: Proceedings of the 2004 IEEE Conference on Probabilistic Methods Applied to Power Systems, Ames, Iowa, US, September 2004, 729–736.

17. Pinson P. and Karniotakis G., *On-line Prediction Risk Assessment for Wind Power Production Forecasts*, Wind Energy, 2004, 7, 119–132.

18. Nielsen H.Aa., Madsen H. and Nielsen T.S., *Using Quantile Regression to Extend an Existing Wind Power Forecasting System with Probabilistic Forecasts*: CD-Proceedings of the 2004 European Wind Energy Conference and Exhibition, London, United Kingdom, November 2004.