A Protocol for Standardazing the performance evaluation of short term wind power prediction models

Henrik Madsen, Pierre Pinson, Georges Kariniotakis, Henrik Aa. Nielsen, Torben Skov Nielsen

Project funded by the European Commission under the 5th (EC) RTD Framework Programme (1998 - 2002) within the thematic programme "Energy, Environment and Sustainable Development"

Project ANEMOS

Contract No.: ENK5-CT-2002-00665

"Development of a Next Generation Wind Resource Forecasting System for the Large-Scale Integration of Onshore and Offshore Wind Farms"

Deliverable 2.3

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

AUTHOR:	H. Madsen
AFFILIATION:	Technical University of Denmark, IMM
ADDRESS:	Ricard Petesens Plads, Bld. 321, 2800 Lyngby, Denmark
EMAIL:	hm@imm.dtu.dk
FURTHER AUTHORS:	G. Kariniotakis, H.Aa. Nielsen, T.S Nielsen, P. Pinson
APPROVER:	H. Madsen

Document Information

DOCUMENT TYPE	Deliverable
DOCUMENT NAME:	ANEMOS_D2.3_EvaluationProtocol.doc
REVISION:	3
REV.DATE:	2004-03-26
CLASSIFICATION:	R0: General Public
STATUS:	Approved

Abstract: In this report, a standardized protocol is proposed for the evaluation of short-term wind power prediction systems. It describes a number of reference prediction models, and it is argued that the use of persistence as a reference prediction models leads to slightly misleading and over-optimistic conclusions about the performance. The use of the protocol is demonstrated using results from both on-shore and off-shore wind farms. The work is a part of the ANEMOS project (EU R&D project), where the protocol is used to evaluate more than 10 prediction systems.

A PROTOCOL FOR STANDARDIZING THE PERFORMANCE EVALUATION OF SHORT-TERM WIND POWER PREDICTION MODELS

H. Madsen[1]*, G. Kariniotakis[2], H.Aa. Nielsen[1], T.S. Nielsen[1], P. Pinson[2]

[1] Technical University of Denmark, Informatics and Mathematical Modeling, DK-2800 Lyngby
[2] Ecole des Mines de Paris, Center for Energy Studies, F-06904 Sophia-Antipolis Cedex

Abstract

In this paper a standardized protocol is proposed for the evaluation of short-term wind power prediction systems. The paper also describes a number of reference prediction models, and it is argued that the use of persistence as a reference prediction models leads to slightly misleading and over-optimistic conclusions about the performance. The use of the protocol is demonstrated using results from both on-shore and off-shore wind farms. The work is a part of the ANEMOS project (EU R&D project) where the protocol is used to evaluate more than 10 prediction systems. Finally, the paper briefly describes the need for future research; in particular in developing more reliable methods for assessing the uncertainty of the predictions, and for evaluating the performance of the uncertainty measures provided by prediction systems.

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

*To whom the correspondance should be addressed (Email: hm@imm.dtu.dk)

1 Introduction

Short-term forecasting of wind energy 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 through the improvement of prediction systems' performance is one of the priorities in wind energy research needs for the coming years (Thor and Weis-Taylor (2003)). Especially, in a liberalized electricity market, prediction tools enhance the position of wind energy compared to other forms of dispatchable generation. Following an emerging demand, there is nowadays an offer for such forecasting tools by various industrial companies or research organizations.

In recent conferences (e.g. Global Windpower 2002, EWEC 2003, etc.) several prediction platforms have been presented (Bailey et al., 1999; Focken et al., 2001; Giebel et al., 2003; Kariniotakis and Mayer, 2002; Landberg and Watson, 1994; Madsen, 1995; Madsen et al., 2000; Marti et al., 2001; Nielsen and Madsen, 1997; Nielsen et al., 2001, 1999). There, an important feedback came from end-users on the necessity to use some standardized methodology when presenting results on the accuracy of a prediction model in order to have a clear idea on the advantages of a specific approach compared to the state-of-the art.

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

This paper presents a complete protocol, consisting of a set of criteria, for the evaluation of a wind power prediction system. This protocol is a result of work performed within the frame of the Anemos Project (22 partners), where the performance of more than 10 prediction systems was evaluated on several on-shore and off-shore case studies. The Anemos project is an R&D project on short-term wind power prediction financed in part by the European Commission, and the project has 22 partners from 7 countries.

To develop this evaluation protocol the criteria found in the bibliography on wind power prediction (around 150 references) were reviewed in detail, and problems with the use of some of the statistics are briefly mentioned. Furthermore, a set of reference predictors is introduced such as persistence, global mean, and a new reference model. Example results are given on a real case study. Finally, guidelines are produced for the use of the criteria.

The aim of this paper is to propose to the scientific community and to end-users a standardized protocol for the evaluation of short-term wind power prediction systems. Nowadays there is an emergence of prediction systems and models (e.g. at the last EWEC03 Conference of Madrid there were more than 50 papers on wind prediction) developed either by research organizations or industrial companies. The choice of such a system is conditioned by the accuracy of the proposed

model. In the bibliography, or in commercial presentations, there is a variety of criteria used to express the accuracy of a model. Recent examples have shown, especially when there is a clear commercial interest, that standard statistical criteria are often not used in the standard way leading to erroneous conclusions on the accuracy of the models.

Based on this the objectives of the paper are:

1. **To present** a proposal for a standardized protocol for evaluating the performance of a model for the short-term prediction of wind power. Moreover, reference predictors will be defined. These predictors are simple models; the performance of which is compared to that of advanced models. By this way, decisions can be taken if it is worthwhile to invest in an advanced model.

2. **To demonstrate** the use of this protocol using results from real case studies.

3. **To present** guidelines on the use of statistical criteria to evaluate the accuracy of a prediction model. Moreover, issues related to evaluating the uncertainty of such models in an on-line environment will be presented.

2 Standard error measures and statistics

In this section, we introduce the notations that are commonly used in the wind power forecasting community. Then, after the presentation of the reference models that may be used as benchmark, the definitions of the usual error measures and statistics will be given. They will form the basis for evaluating the performance of prediction models.

2.1 Notations

P_{inst}	: Wind farm installed capacity
$k = 1, 2, .., k_{max}$	: Prediction horizon (No. of time-steps)
k_{max}	: Maximum prediction horizon
N	: Number of data used for the model evaluation
$P(t+k)$	: Measured power at time $t+k$
$\hat{P}(t+k\|t)$	: Power forecast for time $t+k$ made at time origin t
$e(t+k\|t)$	: Error corresponding to time $t+k$ for the prediction made at time origin t
$\epsilon(t+k\|t)$	: Normalized prediction error (normalized with the installed capacity)

2.2 Reference models

It is worthwhile to develop and implement an advanced wind power forecasting tool if it is able to beat reference models, which are the result of simple considerations and not of modeling efforts. Probably the more common reference model used in the frame of wind power prediction or in the meteorological field is the persistence. This naive predictor states that the future wind generation will be the same as the last measured 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 (saying up to 4-6 hours). This is due to the scale of changes in the atmosphere, which are actually slow. A generalization of the persistence model is to replace the last measured value by the average of the last n measured value

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

Such kind of models is sometimes referred as moving average predictors. Asymptotically (as n goes to infinity), they tend to the global average

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

where $P(t)$ is the average of all the available observations of wind power at time t.

This last one can also be seen as a reference model, but since it is not very dynamic, its performance may be very poor for the first prediction horizons. However, for further look-ahead times, its skill is far better than the one of persistence. The performance of these two reference models has been analytically studied in Nielsen et al. (1998). 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 a_k is defined as the correlation coefficient between $P(t)$ and $P(t+k)$.

All the important statistical quantities, like $\overline{P(t)}$, n and a_k, must be estimated or fixed using the training set, c.f. also the discussion in Section 2.3.

2.3 Training and test data

The generalization performance of a model relates to its prediction capability on new and independent test data. Assessment of this performance is extremely important, since these data gives us a measure of the quality of the prediction model in practice.

It is thus important to evaluate the error measures, which will be proposed in the next section, on data which has not been used for constructing the prediction model or for tuning some parameters of the method. For this reason the data must be split into a training and a test period as illustrated in Figure 1. Some procedures for model building need a validation set for decisions on the models structure – for instance by cross validation. Any such validation data is a part of the training set shown in the figure. Error measures related to the training set are called in-sample measures, while measures related to the test set are called out-of-sample measures.

Unfortunately 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 will perform poorly, and this will be clearly seen from the performance for the test period.

Hence, it is important that the prediction model is developed and tuned based on the training data without considering the test data. Hereafter the model obtained should be applied to the test data, mimicking the actual application, and the error measures reported should be based on the test period only.

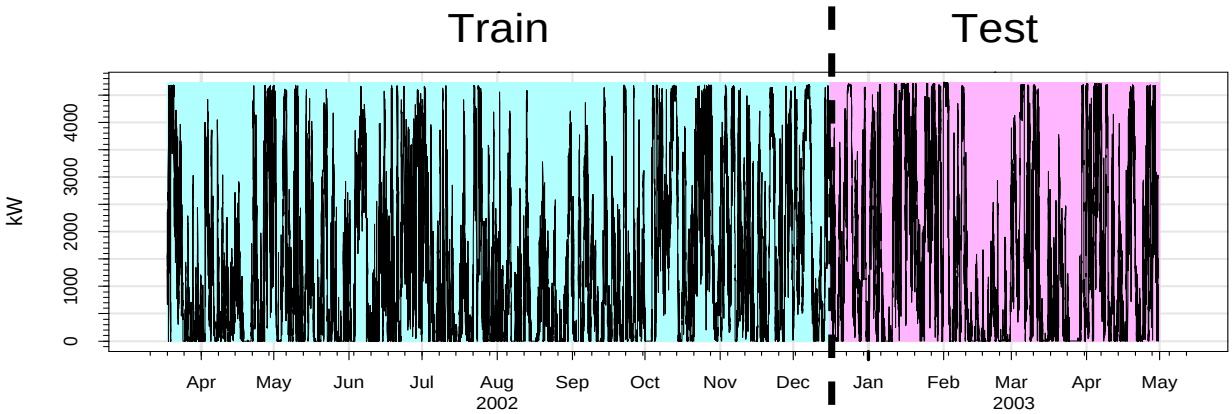

Figure 1: A DATA SET FROM THE OFF-SHORE WIND FARM TUNØ KNOB IN DENMARK SPLIT INTO A TRAINING AND A TEST PERIOD.

2.4 Definition of error measures

Prediction error definitions

In the field of time series prediction in general, the prediction error is defined as the difference between the measured and the predicted value. Therefore, since we consider separately each forecast

horizon, *the prediction error* for the lead time k is defined as

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

Very often it is convenient to introduce the *normalized prediction error*

$$\epsilon(t+k|t) \;=\; \frac{1}{P_{inst}}\big(P(t+k) - \hat{P}(t+k|t)\big), \tag{6}$$

where P_{inst} is the installed capacity.

Let p denote the number of estimated parameters using the considered data. Hence for the test data $p = 0$. In the following N is the number of prediction errors.

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

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

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

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

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

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

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

Notice, that both systematic and random errors contribute to the MAE-value.

Before introducing the $RMSE$ it is useful to introduce the Mean Squared Error

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

The Root Mean Squared Error is then simply

$$RMSE(k) = \sqrt{MSE} \tag{11}$$

$$= \sqrt{\frac{\sum_{t=1}^{N}(e(t+k|t))^2}{N-p}}. \tag{12}$$

Both systematic and random errors contribute to the $RMSE$ criterion.

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

$$SDE(k) = \Big(\frac{\sum_{t=1}^{N}(e(t+k|t) - \overline{e(k)})^2}{N-(p+1)}\Big)^{\frac{1}{2}}. \tag{13}$$

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 $BIAS(k)$ and $MAE(k)$ are associated with the first moment of the prediction error, and hence these are measures which are directly related to the produced energy. The values of $RMSE(k)$ and $STD(k)$ are associated with the second order moment, and hence to the variance of the prediction error. For the 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 $\epsilon(t+k|t)$. The interest of using normalized error measures is to produce results independent of wind farm sizes.

Some references use other definitions of error measures. One example is the so-called surplus for given period, which is the sum of all positive prediction errors.

Comparison of models

It might be of interest to highlight and to quantify the gain of preferring an advanced approach to the reference ones. This gain, denoted as an improvement with respect to the considered reference model, is

$$Imp_{ref,EC}(k) = \frac{EC_{ref}(k) - EC(k)}{EC_{ref}(k)}, \tag{14}$$

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

An 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 average model (cf. Equation (3)) where the average is estimated for 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 to avoid the use of this criterion as a main tool for performance evaluations in general. If, for instance the naive prediction is used for large horizons, the resulting R^2-value will be negative! This is due to the fact that the asymptotic variance of the prediction errors for the naive prediction is twice the variance of the global mean prediction defined by Equation (3), cf. Nielsen et al. (1998). 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.

There exists several possible definitions of the R^2-value. One frequently used possibility is to define the R^2-value using the correlation between the measured and predicted wind power. The problem of this definition is that even though the predictions might be biased (and/or relative biased) this definition will lead to $R^2 = 1$. The above suggested definition 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

Obviously, the general performance of the prediction method influences the value of the error measures. However, the site and period may also significantly influence the apparent performance of a given forecasting system. Figure 2 shows results obtained with the same prediction method for five different sites/periods. From the plot, one can notice that for Klim and Tunø, which are both located in Denmark, the model performance differs by approximately 20% (2 percent point). This may both be due to an effect of site, and the fact that the period is different for the two sites.

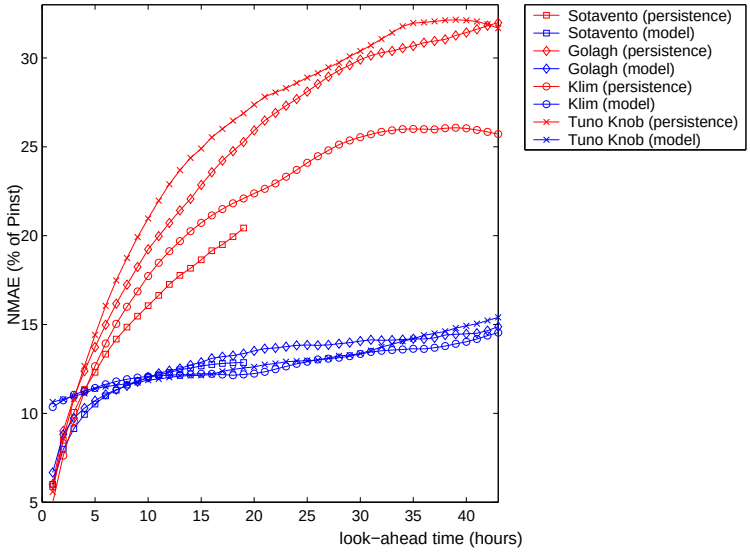

Figure 2: PERFORMANCE IN TERMS OF NMAE OF TWO PREDICTORS (PERSISTENCE AND A STATE-OF-ART STATISTICAL PREDICTION METHOD) FOR FOUR SITES/PERIODS (GOLAGH, KLIM, SOTAVENTO, AND TUNØ KNOB).

3 Exploratory analysis

There exist a large number of other tools for exploratory analysis, and some of the methods which are found to be of particular interest in relation to wind power prediction will be illustrated. These tools for exploratory analysis of the prediction errors provide a deeper insight into the performance of the methods.

A histogram plot showing the distribution of prediction errors is very useful. 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, like high wind speed, summer, westerly wind, etc. An example of using the histogram will be shown for the case study considered in Section 4.

Another useful tool is to plot the cumulated squared prediction errors. Figure 3 shows the cumulated squared errors for 6 hour predictions for the Tunø off-shore wind farm. The cumulated plot shows a clear change in the increment for the cumulated squared prediction errors for the last couple of weeks of the considered period, and this should then lead to further investigations.

The use of the six hour horizon in the cumulated squared prediction errors is found to be useful for detecting changes in the numerical weather predictions.

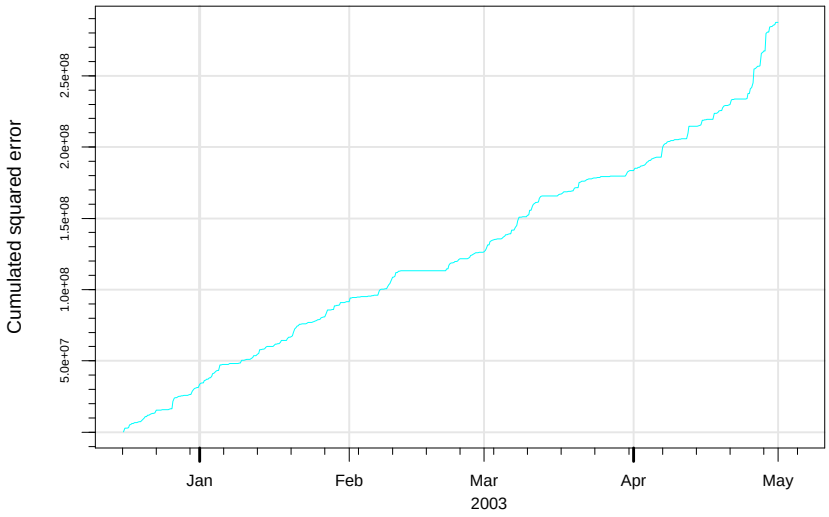

Figure 3: CUMULATED SQUARED PREDICTION ERRORS FOR THE TUNØ OFF-SHORE WIND FARM

4 Application to a real case study

As an illustration of the previously described error measures and statistics, we consider the case study of a real multi-MW wind farm located in Ireland. A state-of-the-art statistical prediction model is used for giving two-day ahead estimations of the wind farm hourly power generation, with Hirlam Numerical Weather Predictions (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 wind power forecasting model is evaluated over a 3-month period corresponding approximately to winter 2003.

Figure 4 shows the prediction model normalized bias ($NBIAS$) as a function of the lead time, showing values between -0.14% and 0.01%. Actually, this means that for this case study, the model does not make a systematic error. This is a nice property that is wanted when using a prediction model. Nowadays, both statistical models 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 (the persistence is used here), for every prediction horizon. The $NMAE$ can be interpreted directly: for instance, the advanced approach experienced an average error representing 13% of the installed power for its one-day ahead predictions, over the whole evaluation period. Such an information is

not provided by the $NRMSE$, since it considers squared errors. The $NRMSE$ measure is most relevant if small errors are of minor importance compared to large prediction errors.

The model skills are then compared by calculating the error reduction that allows the model with respect to the reference one. An advanced prediction approach should propose a significant gain over the reference models, in order to justify the modeling efforts involved in their design. Here, the improvement owing to the model ranges from -10% for the first look-ahead time to almost 55% for longer-term predictions (for both criteria). Beating the persistence for the first horizons is not easy, although for longer-term (12-48 hour ahead) very large improvements can be achieved. This is why the new reference model introduced above, which is the best reference competitor over the whole horizon range, should be considered instead of the persistence.

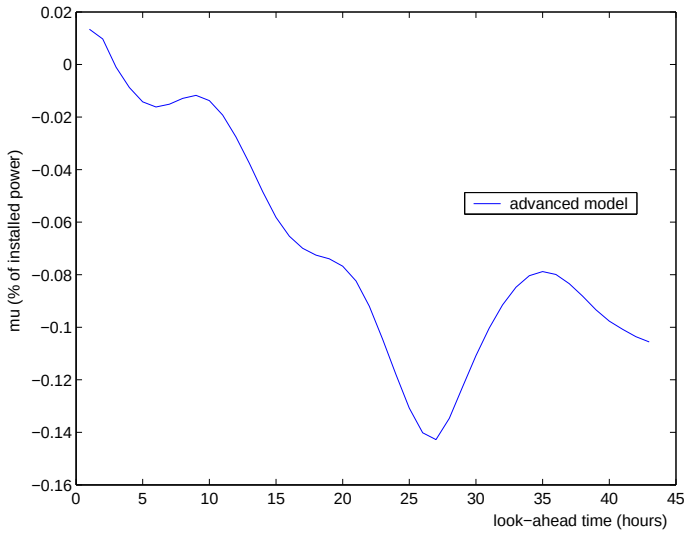

Figure 4: PREDICTION MODEL BIAS AS A FUNCTION OF THE LEAD TIME.

Finally, more subtle information can be extracted from error distributions as shown in Figure 6. They are produced for the 1^{st} and 24^{th} lead times, with bins representing 5% of the rated power. A first glance at the histogram sharpness, skewness, inf/sup bounds, already gives a first idea on the model performance. Comparing the two histograms of Figure 6, one can notice that the error distributions are almost perfectly symmetric and centered around 0, and that the one for one-hour ahead predictions is a lot sharper than the other. During the evaluation periods, the model never experienced errors greater than 40% of P_{inst} for the first lead time; this is not the case for 24-hour ahead forecasts. The optimal number of bins used in the histogram is related to the range of the data (range(x)) and the number of samples, N. In Scott (1992) the suggested optimal range, w, for a single bin is

$$w = \text{range}(x)/(\log_2(N) + 1). \tag{16}$$

It is recommended to use the same size for all bins (i.e. 5% for the case of Figure 6) when plotting

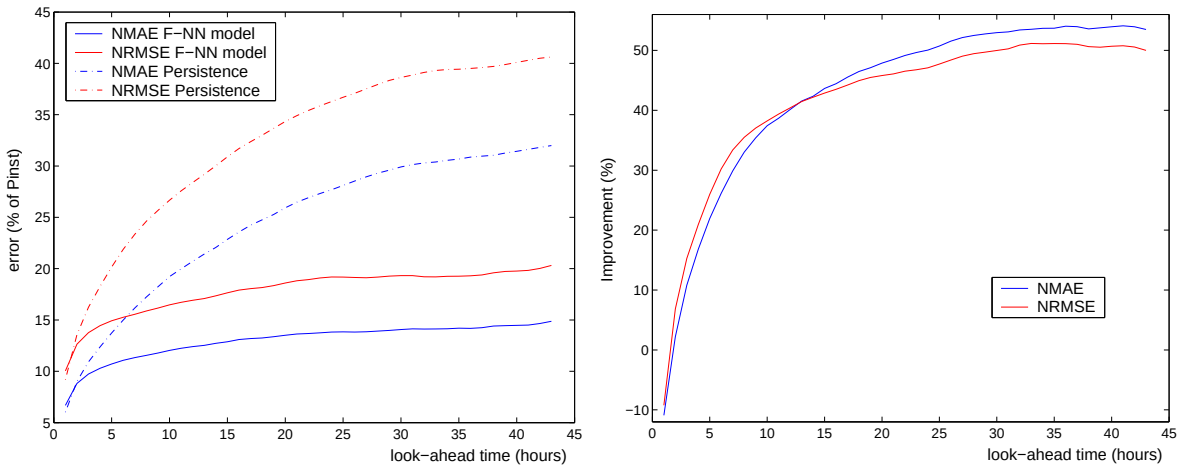

Figure 5: USE OF THE $NMAE$ AND THE $NRMSE$ FOR ASSESSING THE PERFORMANCE OF THE ADVANCED PREDICTION APPROACH, AND FOR COMPARISON WITH ONE OF THE REFERENCE PREDICTORS.

an histogram to avoid misleading interpretations of the error distributions.

Moreover, this classification of the errors allows one to highlight statistics about the frequency of occurrence of errors below or above a certain level. For instance, the prediction model errors for this case-study are:

- less than 7.5% of the wind farm nominal power 68% of the times for the first lead time,
- less than 7.5% of P_{inst} 24% of the times for lead time 24,
- higher than 17.5% of P_{inst} 3% of the times for the first horizon, etc.

The combination of all these error measures and statistics gives a useful global view on a prediction model skills for end-users interested in assessing the performance of the forecasting tool they use, and comparing such a performance for different models and/or for different sites. However, this thorough evaluation has also a great interest for people involved in the research and design of wind power prediction methods. Indeed, a detailed understanding of the prediction error characteristics is needed for proposing future improvements of the methods.

5 Guidelines and recommendations

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

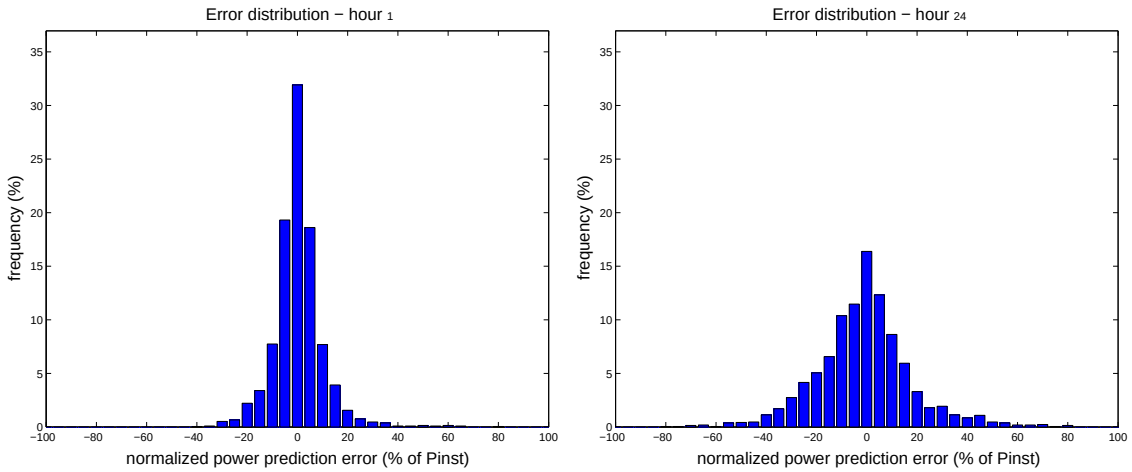

Figure 6: NORMALIZED PREDICTION ERROR DISTRIBUTIONS FOR THE FIRST LOOK-AHEAD TIME (LEFT) AND FOR LEAD TIME 24 (RIGHT).

5.1 Recommendations

Regarding the performance measures we have the following recommendations:

- 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 the quality of the considered data (i.e. detection of missing or erroneous data) should be performed before to start with the performance evaluation.
- As a minimum set of error measures use:
  - 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 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 also for the selected simple reference models.

Finally, it should be realized that the most appropriate measure depends on the intended application.

5.2 Operational framework

Before presenting any performance measure it is very important to specify the operational framework.

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).
- Sampling strategy. Specify whether the data 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.
- Frequency of updates. Actually, some models only give forecasts when NWPs are provided (i.e. every 6, 12 or 24 hours) when some others operate with a sliding window (typically one hour) since they consider on-line production data as input.
- Characteristics of NWP forecasts (frequency of delivery, delay in delivery, horizon, time step, resolution, grid values or interpolated at the position of the farm).
- Use of SCADA data as input. Specify which SCADA data is used, and the sampling strategy for the data.

In the description above we have focused on a single wind farm. The modifications needed for considering the wind power predictions for larger areas are minor given that the relevant data is available.

6 Conclusion and discussion

There is a large need for standardizing the error measures and the reference models for characterizing the performance of models for wind power prediction. Use of comparisons with the persistence predictor does not give a fair measure of the performance of the model, since even the use of the

long-term average as the prediction leads to a reduction of 50% in the variance of the prediction error compared to the prediction error obtained by persistence Nielsen et al. (1998).

In this paper guidelines for evaluating wind power predictions are presented, and a minimum set of suggested error measures is described. For performance comparisons it is important not only to use proper performance measures, but also to use the same data. It is very important to use test data, and not the data used for estimating/training the model, for comparisons.

Besides the limited number of recommended error measures the researcher should perform further (exploratory) analyses of the prediction errors; comparisons with other (simple) predictors, histograms, plots of cumulated squared errors, etc. This allows a deeper understanding of the limitations of the method and points towards improvements.

The presented measures are mostly designed for off-line evaluations. Some of the measure might also be used in on-line situations. Still more and more methods are established for providing also the uncertainty of the prediction. In the latter part of the Anemos project we will elaborate on performance measures which focus on an evaluation of the provided uncertainty, and this will be a subject of increasing interest for future research programs dealing with on-line wind power predictions.

The sequence of prediction errors is obviously correlated, and the so-called autocorrelation of this time series might be of importance for the user. That holds in particular for users having some sort of energy storage. Hence, an operational approach for presenting the autocorrelation of the error sequence is needed, and this subject will also be dealt with in a later stage of the Anemos project.

Acknowledgements

This work has been supported by the European Commission under the 5th Framework Program, contract ENK5-CT-2002-00665, as part of the ANEMOS project. Anemos is a EU R&D Project aiming to develop accurate models for on-shore and off-shore wind resource forecasting using statistical as well as physical methods. As part of the project an integrated software system, Anemos, will be developed to host the various models. This system will be installed by several utilities for on-line operation at on-shore and off-shore wind farms for local as well as regional wind power prediction.

References

B. Bailey, M.C. Brower, and J. Zack. Short-term wind forecasting: development and application of a mesoscale model. In *European Wind Energy Conference*, pages 1062–1065, 1999.

U. Focken, M. Lange, and H.-P. Waldl. Previento - a wind power prediction system with an innovative upscaling algorithm. In *European Wind Energy Conference*, pages 826–829, 2001.

G. Giebel, G. Kariniotakis, and R. Brownsword. The state of the art in short-term prediction of wind power. Technical report, http://anemos.cma.fr, 2003. Deliverable report of the EU project ANEMOS.

G. Kariniotakis and D. Mayer. An advanced on-line wind resource prediction system for the optimal management of wind parks. In *Proceedings of 2002 Global Windpower*, Paris, France, April 2002.

L. Landberg and S.J. Watson. Short-term prediction of local wind conditions. *Boundary-Layer Meteorol.*, 70, 1994.

H. Madsen, editor. *Wind Power Prediction Tool (WPPT) in Central Dispatch Centres*. ELSAM, Fredericia, Denmark, 1995.

H. Madsen, T.S. Nielsen, H.Aa. Nielsen, and L. Landberg. Short-term prediction of wind farm electricity production. In *European Congress on Computational Methods in Applied Sciences and Engineering*, Barcelona, 11–14 Sept. 2000.

I. Marti, T. S. Nielsen, H. Madsen, J. Navarro, and C. G. Barquero. Prediction models in complex terrain. In *Proceedings of the European Wind Energy Conference*, Copenhagen, Denmark, 2001.

T.S. Nielsen, A. Joensen, H. Madsen, L. Landberg, and G. Giebel. A new reference model for wind power forecasting. *Wind Energy*, 1:29–34, 1998.

T.S. Nielsen and H. Madsen. Statistical methods for predicting wind power. In *Proceedings of the European Wind Energy Conference*, pages 755–758, Dublin, Eire, 1997. Irish Wind Energy Association.

T.S. Nielsen, H. Madsen, H.Aa. Nielsen, L. Landberg, and G. Giebel. Zephyr - the prediction models. In *European Wind Energy Conference*, pages 868–871, 2001.

T.S. Nielsen, H. Madsen, and J. Tøfting. Experiences with statistical methods for wind power prediction. In *Proceedings of the European Wind Energy Conference*, pages 1066–1069, Nice, France, 1999. James & James (Science Publishers).

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

S.-E. Thor and P. Weis-Taylor. Long-term research and development needs for wind energy for the time frame 2000-2020. *Wind Energy*, 5:73–77, 2003.