Methods for the Estimation of the Uncertainty of Wind Power Forecasts

P. Pinson*, **H.Aa. Nielsen, H. Madsen**
Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark
M. Lange
Energy and Meteo Systems, Oldenburg, Germany
G. Kariniotakis
Center for Energy and Processes, Ecole des Mines de Paris, France

March 23, 2007

Abstract

This report describes the scientific developments carried out in the frame of the Work Package 3 of the ANEMOS project. It concentrates on the developed methods, specially dedicated to the estimation of the uncertainty of wind power forecasts. Several methods for quantile or interval forecasting are described, and their performance is evaluated and discussed. Focus is particularly given to the fact that the uncertainty in wind power prediction is highly conditional to a wealth of explanatory variables. Among others, the meteorological situations and the nonlinear nature of the power curve greatly impact the forecast uncertainty. Specific developments towards meteorological-situation specific uncertainty estimates are also given.

Key words:

ANEMOS project, wind power, forecasting, uncertainy, prediction interval, quantile forecast, quantile regression, adapted resampling, meteorological situations, risk indices.

* *Corresponding author:*
P. Pinson, Informatics and Mathematical Modelling, Technical University of Denmark,
Richard Petersens Plads (bg. 321 - 020), DK-2900 Kgs. Lyngby, Denmark.
Tel: +45 4525 3428, fax: +45 4588 2673, email: pp@imm.dtu.dk, webpage: www.imm.dtu.dk/∼pp

Contents

Chapter 1

Introduction

Predictions of wind power output are traditionally provided in the form of point forecasts. They have the advantage of being easily understandable because this single number is expected to tell everything about future power generation. Today, a major part of the research efforts on wind power forecasting still focuses on point prediction[1] only, with the aim of assimilating more and more observations in the models or refining the resolution of physical models for better representing wind fields at the very local scale for instance [32]. These efforts may lead to a significant decrease of the level of prediction error. However, even by better understanding and modeling both the meteorological and power conversion processes, there will always be an inherent and irreducible uncertainty in every prediction. This epistemic uncertainty corresponds to the incomplete knowledge one has of the processes that influence future events [87].

Therefore, in complement to point forecasts of wind generation in the next hours, of major importance is to provide means for assessing online the accuracy of these predictions. Error measures described in the Anemos Deliverable Report 2.1 [64] only provide an assessment of a given point forecasting method performance over a large period of time. They tell what is the historic performance of the method, but they cannot give an estimation of the uncertainty related to a given prediction. In practice today, uncertainty can be expressed:

- with indices informing on the predictability of individual weather situations,
- with probabilistic predictions in the form of quantile of interval forecasts. This last type of uncertainty estimates are associated to a probability related to the likeliness of future power production. If not, they can be seen as error bands, giving a qualitative and visual information on expected uncertainty.

Whatever their nature, such uncertainty estimates are expected to be valuable for developing alternative strategies for the management or the trading of wind power generation. In a general manner, they are necessary for optimizing the decision-making process related to the use of wind power forecasts.

[1] Point prediction is defined as the providing of a single forecast power value for each look-ahead time for the considered site (or group of sites), thus without adressing the issue of uncertainty.

The present report gathers the research works carried out in the frame of the European project Anemos for the estimation of the uncertainty of wind power forecasts. More precisely, the proposed methods have been developed by Energy & Meteo Systems (and University of Oldenburg), Technical University of Denmark, and ARMINES/École des Mines de Paris. The various approaches have been derived either from studies on the predictability of the weather dynamics, or from physical considerations on the wind-to-power conversion process, or finally from statistical methods (based on a non-parametric modeling of predictive distributions). Overall conclusions are given at the end of the present report, as well as perspectives regarding further research.

Chapter 2

Assessment of wind speed dependent prediction error

Abstract

The investigations in this chapter follow the idea that the prediction error quantitatively depends on the meteorological situation that has to be predicted. As a first approach the wind speed as a main indicator of the forecast situation is considered in greater detail. The probability density functions (pdf) of the measured wind speed conditioned on the predicted one are found to be Gaussian in the range of wind speeds that is relevant for wind energy applications. An analysis of the standard deviations of these conditional pdfs reveals no systematic dependence of the accuracy of the wind speed prediction on the magnitude of the wind speed. With the pdfs of the wind speed as basic elements the strongly non Gaussian distribution of the power prediction error is explained underlining the central role of the non linear power curve. Moreover, the power error distribution can easily be estimated based on the statistics of the wind speed, the wind speed forecast error, and the power curve of the turbine. Thus, it can be reconstructed without knowing the actual measured power output which is interesting for future sites or sites where no data is available. In addition, a simple formula based on linearising the standard deviation of the error is derived. This model illustrates the dominating effect of small relative errors in the wind speed prediction being amplified by the local derivative of the nonlinear power curve. This chapter is taken from [57].

2.1 Idea behind detailed error assessment

The standard error measures that are used for the performance evaluation of prediction methods are based on annual averages of the data and provide only one constant value for each forecast time. However, there is reason to believe that the magnitude of the error quantitatively depends on the meteorological situation to be predicted. Thus, a more detailed view on the prediction error is required where the main parameters that characterise typical wind conditions have to be identified and related to the corresponding forecast error.

The first candidate that can serve as an indicator of the forecast situation is the wind speed itself. It is a continuous parameter closely related to the forecast situation and the main input into the power prediction system such that in this chapter the role of the

predicted wind speed and its prediction error are investigated more deeply.

The chain of events seems rather straightforward: the initial uncertainty introduced by the error of the wind speed prediction is propagated through the power prediction system where it is mainly subject to changes by the power curve. Due to its non linearity the power curve is expected to amplify or damp initial errors in the wind speed according to its local derivative. And this derivative is a function of the wind speed. So clearly, the power curve is expected to be the key element that connects the errors of the wind speed prediction and the power prediction. Thus, the major aim here is to derive a quantitative relation between the two.

The practical use of this relation and its implementation into a wind power prediction system is to provide the user with additional information to estimate the risk of trusting in the prediction. Hence, the prediction system has to supply the prediction itself and a useful indication concerning the reliability of the individual prediction. As financial losses might be proportional to the magnitude of the prediction error the inherent risk of faulty predictions must be known for each forecast situation.

2.2 Introduction of conditional probability density functions

It was seen in previous investigations [53, 57] that the probability density function, $\mathrm{pdf}(\epsilon_u)$, of wind speed error is Gaussian in most cases. As mentioned above a more detailed information concerning the error to be expected in special conditions is desirable. Thus, a first approach is to refine the pdfs and look at the statistical properties of the measured wind when the predicted wind speed is confined to a certain value which from a mathematical point of view leads to conditional pdfs. In this section the deviations of the measured wind speed from one is investigated in terms of these conditional pdfs.

Predicted and measured values of a meteorological variable at the same point of time and space are naturally not independent. In fact, they are supposed to be highly correlated as this is a major prerequisite for an accurate prediction. The pairs $(x_{\mathrm{pred}}, x_{\mathrm{meas}})$ are drawn simultaneously from a joint distribution, $\mathrm{pdf}(x_{\mathrm{pred}}, x_{\mathrm{meas}})$, that characterises the statistical properties of the prediction and its error. This means that for arbitrary but fixed predicted values x_{pred} the occurrences of the corresponding measured values x_{meas} are expected to be mainly concentrated in an interval around x_{pred} rather than being spread over the whole range of all possible values.

In the following investigation the "prediction perspective" is taken which means that the predicted values are used as condition to the measurements. This aims at formulating the equations such that they can directly be used for prediction purposes.

Formally the conditional pdf of the wind speed is given by

$$\mathrm{pdf}(u_{\mathrm{meas}}|u_{\mathrm{pred}}) = \frac{\mathrm{pdf}(u_{\mathrm{pred}}, u_{\mathrm{meas}})}{\mathrm{pdf}(u_{\mathrm{pred}})} \tag{2.1}$$

where $\mathrm{pdf}(u_{\mathrm{pred}}, u_{\mathrm{meas}})$ is the joint distribution and $\mathrm{pdf}(u_{\mathrm{pred}})$ the unconditional, so-called marginal, wind speed distribution.

For practical purposes the time series are given by a finite number of data points with a certain accuracy. So the probability function of the measured wind speed u_{meas} under

the condition u_{pred} is approximately calculated by confining the prediction values to an interval around u_{pred} and bin-counting the corresponding values for u_{meas}.

The conditional pdfs can, of course, be used to obtain the unconditional distribution by

$$\text{pdf}(u_{\text{meas}}) = \int_0^\infty \text{pdf}(u_{\text{meas}}|u_{\text{pred}})\,\text{pdf}(u_{\text{pred}})\,du_{\text{pred}}. \tag{2.2}$$

These conditional pdfs of the wind speed can be used to reconstruct the error distribution, pdf(ϵ_u), of the wind speed prediction. The conditional pdfs defined in Equation (2.1) already contain all the information needed for that as they describe the deviations between predicted and measured values. With the transformation

$$u_{\text{meas}} \mapsto \epsilon_u \quad \text{where} \quad \epsilon_u = u_{\text{pred}} - u_{\text{meas}} \tag{2.3}$$

the pdf($u_{\text{meas}}|u_{\text{pred}}$) are merely mirrored and shifted along the abscissa and transfered to pdf($\epsilon_u|u_{\text{pred}}$). The original, unconditional error distribution can then be recovered by

$$\text{pdf}(\epsilon_u) = \int_0^\infty \text{pdf}(\epsilon_u|u_{\text{pred}})\,\text{pdf}(u_{\text{pred}})\,du_{\text{pred}}. \tag{2.4}$$

Note that if the conditional pdfs are approximately equal to each other and do not vary much with u_{pred} one obtains $\text{pdf}(\epsilon_u) \approx \text{pdf}(\epsilon_u|u_{\text{pred}})$. In this case the overall pdfs of the error can directly serve as conditional pdf.

Reconstructing the distribution of the power prediction error

Knowing the statistical properties of the wind speed is very helpful with regard to the prediction of the power output as the wind speed is the main input variable fed into the power prediction system. However, the error distributions of the power prediction differ qualitatively from those of the wind speed in that they are neither Gaussian nor approximately Gaussian which is mainly due to the non linear power curve of the wind turbines. The following investigation focuses on the role of the power curve in transforming the distributions of the wind speed and its error into those of the power output.

In Figure 2.1 a typical power curve, $P(u)$, is shown with cut-in speed around $4\ m/s$ followed by a sharp increase of power over the interval $5 - 10\ m/s$ and a saturation at the level of rated power for higher wind speeds. Let Δu be a small interval around the wind speed u and $\Delta P = P(u + \Delta u) - P(u)$ the resulting difference in the power output. For small Δu the corresponding ΔP is then given by a Taylor expansion around u:

$$\Delta P = \frac{dP}{du}(u)\Delta u. \tag{2.5}$$

This equation generally describes how wind speed intervals are mapped to power intervals. If Δu is regarded as a small deviation between predicted and measured wind speed Equation (2.5) illustrates that the power curve scales errors in the wind speed according to its local derivative. Thus, whether deviations in the wind speed are amplified or damped depends on the magnitude of the wind speed.

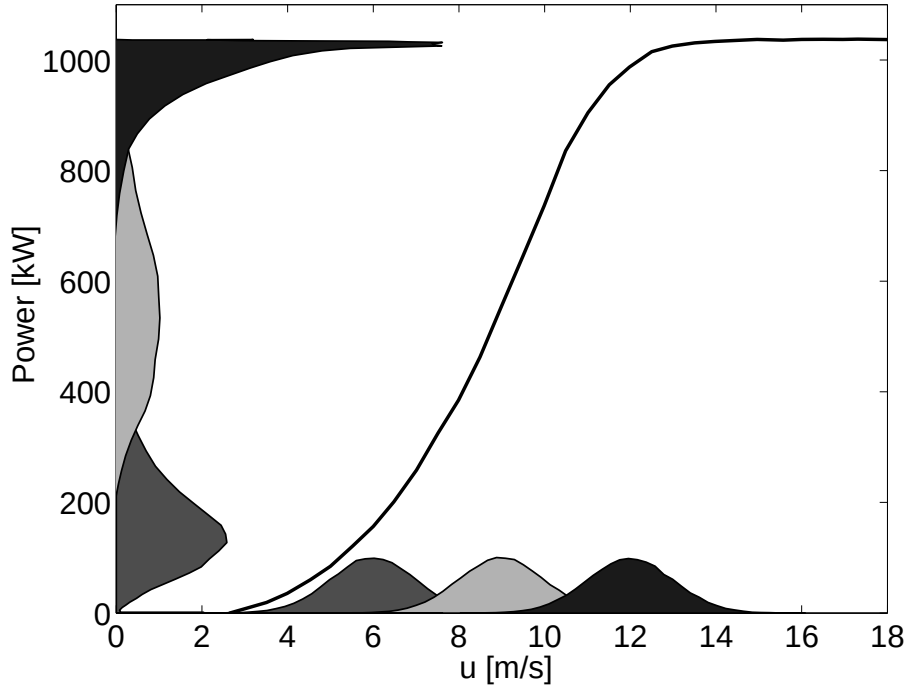

Figure 2.1: *Power curve of a pitch regulated wind turbine (solid line). For three different wind speeds conditional pdfs, pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$), are illustrated on the x-axis which have mean=$u_{\mathrm{pred,i}}$ and standard deviation 1. The corresponding pdf($P(u_{\mathrm{meas}})|P(u_{\mathrm{pred,i}})$) constructed from Equation (2.5) are plotted on the y-axis. The pdfs are not normalised for better visualisation. For small and large wind speeds the Gaussian wind speed distributions are strongly deformed and no longer symmetric. For medium wind speeds the pdf of the power is significantly flatter and broader than the pdf of the wind speeds.*

Using Equation (2.1) the probability to find a measurement value u in the interval $[u_{\mathrm{meas}}, u_{\mathrm{meas}} + \Delta u]$ with u_{pred} confined to $[u_{\mathrm{pred}}, u_{\mathrm{pred}} + \Delta u]$ is

$$w := \mathrm{pdf}(u_{\mathrm{meas}}|u_{\mathrm{pred}})\,\mathrm{pdf}(u_{\mathrm{pred}})\,(\Delta u)^2. \tag{2.6}$$

If the area $(\Delta u)^2$ around $(u_{\mathrm{meas}}, u_{\mathrm{pred}})$ is mapped to $(\Delta P)^2$ around $(P(u_{\mathrm{meas}}), P(u_{\mathrm{pred}}))$ according to Equation (2.5) the probability w is preserved because all events that are recorded in the wind speed intervals also occur in the power output intervals. Hence,

$$\begin{aligned} w &= \mathrm{pdf}(u_{\mathrm{meas}}|u_{\mathrm{pred}})\,\mathrm{pdf}(u_{\mathrm{pred}})\,(\Delta u)^2 \\ &\stackrel{!}{=} \mathrm{pdf}(P(u_{\mathrm{meas}})|P(u_{\mathrm{pred}}))\,\mathrm{pdf}(P(u_{\mathrm{pred}}))\,(\Delta P)^2 \\ &= \mathrm{pdf}(P(u_{\mathrm{meas}})|P(u_{\mathrm{pred}}))\,\mathrm{pdf}(P(u_{\mathrm{pred}}))\left(\frac{dP}{du}(u_{\mathrm{meas}})\Delta u\right)\left(\frac{dP}{du}(u_{\mathrm{pred}})\Delta u\right). \end{aligned} \tag{2.7}$$

If Equation (2.7) is solved for the desired pdfs of the power output one obtains

$$\begin{aligned}&\mathrm{pdf}(P(u_{\mathrm{meas}})|P(u_{\mathrm{pred}}))\,\mathrm{pdf}(P(u_{\mathrm{pred}}))\\ =&\mathrm{pdf}(u_{\mathrm{meas}}|u_{\mathrm{pred}})\,\mathrm{pdf}(u_{\mathrm{pred}})\left(\frac{dP}{du}(u_{\mathrm{meas}})\right)^{-1}\left(\frac{dP}{du}(u_{\mathrm{pred}})\right)^{-1}.\end{aligned} \tag{2.8}$$

This equation provides the essential relation between the distributions of wind speed on the one hand and power on the other hand. As expected the power curve plays a crucial role in connecting the statistical properties of both quantities. Note that the divergence in Equation (2.8) for $(dP/du)^{-1} \to \infty$ is compensated by $\Delta P \to 0$ according to Equation (2.5) such that the probability w is always well defined.

Figure 2.1 illustrates the effect of Equation (2.5) for three different Gaussian wind speed distributions with same standard deviations. For small and large wind speeds the Gaussian wind speed distributions are strongly deformed and no longer symmetric. For medium wind speeds the pdf of the power is significantly flatter and broader than the pdf of the wind speeds.

The next steps towards a full description of the statistics of the power prediction error in terms of the wind speed distributions are now right ahead: Equation (2.8) can be used to create the functions $\mathrm{pdf}(P(u_{\mathrm{meas}})|P(u_{\mathrm{pred}}))$ for each $P(u_{\mathrm{pred}})$. Then these functions are weighted according to the frequency distribution of $P(u_{\mathrm{pred}})$ and their variables are shifted analogous to transformation (2.3). Finally, these shifted functions are added and the unconditional pdf of the power prediction error is reconstructed.

Equations (2.7) and (2.8) suggest that the basic elements in this reconstruction procedure can be defined by

$$F(P(u_{\mathrm{meas}}), P(u_{\mathrm{pred}})) := \mathrm{pdf}(u_{\mathrm{meas}}|u_{\mathrm{pred}})\,\mathrm{pdf}(u_{\mathrm{pred}})\left(\frac{dP}{du}(u_{\mathrm{meas}})\right)^{-1}\left(\frac{dP}{du}(u_{\mathrm{pred}})\right)^{-1} \tag{2.9}$$

Note that this definition is an intermediate step that conveniently summarises all mathematical terms involved. The functions F contain the information how the conditional pdf of the wind speed has to be transformed to the corresponding power pdf and which weight this pdf has. Once the right-hand side of Equation (2.9) has been used F is defined in the power domain, i.e. the independent variables are $\tilde{P}_{\mathrm{meas}} := P(u_{\mathrm{meas}})$ and $P_{\mathrm{pred}} := P(u_{\mathrm{pred}})$.

The notation $\tilde{P}_{\mathrm{meas}}$ denotes the measured power obtained by plugging the measured wind speed at hub height into the theoretical power curve. Generally, $\tilde{P}_{\mathrm{meas}}$ slightly differs from the direct measurement of the power output, P_{meas}, because the power curve of the local wind turbine might deviate from the theoretical curve or additional errors that are not covered by the power curve come in.

Analogous to the transformation of the conditional wind speed distribution in Equation (2.3) the first variable in each of the functions F is shifted according to

$$P_{\mathrm{meas}} \mapsto \epsilon_P \quad \text{where} \quad \epsilon_P = P_{\mathrm{pred}} - \tilde{P}_{\mathrm{meas}}. \tag{2.10}$$

In the final step the unconditional distribution of the power prediction error is obtained by the integration

$$\mathrm{pdf}_{\mathrm{rec}}(\epsilon_P) = \int_0^\infty F(\epsilon_P, P_{\mathrm{pred}})\, dP_{\mathrm{pred}}. \tag{2.11}$$

Despite having a slight touch of looking complicated this method has some benefits. It exclusively provides the error statistics of the power prediction based on three ingredients: the conditional distributions of the wind speed prediction, the derivative of the power curve and the distribution of the predicted wind speeds. For practical purposes these three components are either given or can be estimated. The power curve is typically known numerically such that the derivative can easily be obtained. The distribution of the predicted wind speed is normally provided by the NWP output but if not it can be estimated from prediction data of nearby sites or from Weibull distributions of the measurement data. If the conditional pdfs of the wind speed are not known for the site in question it can be assumed that they are Gaussian with mean values and standard deviations obtained by a qualified guess.

The concepts developed in the this section are now applied to data from real sites to check if the relations between the wind speed and the power distributions can be confirmed with finite sets of data points.

2.3 Conditional PDF of wind speed data

The conditional pdfs of the wind speed are calculated as described by Equation (2.1), i.e. the distribution of the measured values is determined with the corresponding predicted value confined to a certain interval. The range of the occurring u_{pred} is divided into equidistant bins with width Δu which is typically set to 1 m/s. The boundaries of the bins are given by $[u_{\mathrm{pred,i}} - \Delta u/2, u_{\mathrm{pred,i}} + \Delta u/2]$. Hence, $u_{\mathrm{pred,i}}$ denotes the middle of the bins.

In Figures 2.2 and 2.3 the conditional pdfs of the wind speed at two sites with different average wind speeds are shown based on measured and predicted data of the year 1996. In these examples the 36 h values of prediction and measurement are used but the qualitative behaviour of the other prediction times is comparable.

The distributions for small wind speeds are unsymmetric and far from being normal. This is due to the fact that the wind speed is always positive and deviations from $u_{\mathrm{pred,i}}$ are limited towards lower wind speeds but not towards larger ones. With increasing $u_{\mathrm{pred,i}}$ the pdfs become rather symmetric.

Compared to a normal distribution with same mean and standard deviation these pdfs can be considered as being approximately Gaussian although the number of data points available per pdf is relatively small. Two statistical tests are used to systematically check the normality of the conditional pdfs. As expected the hypothesis that the pdf is Gaussian is rejected by the χ^2-test and the Lilliefors-test for distribution functions with small $u_{\mathrm{pred,i}}$. For larger prediction values both tests indicate normality: in Figure 2.2 for the pdfs with $2.5\ m/s \leq u_{\mathrm{pred,i}} \leq 6.5\ m/s$ and in Figure 2.3 for $2.5\ m/s \leq u_{\mathrm{pred,i}} \leq 8.5\ m/s$.

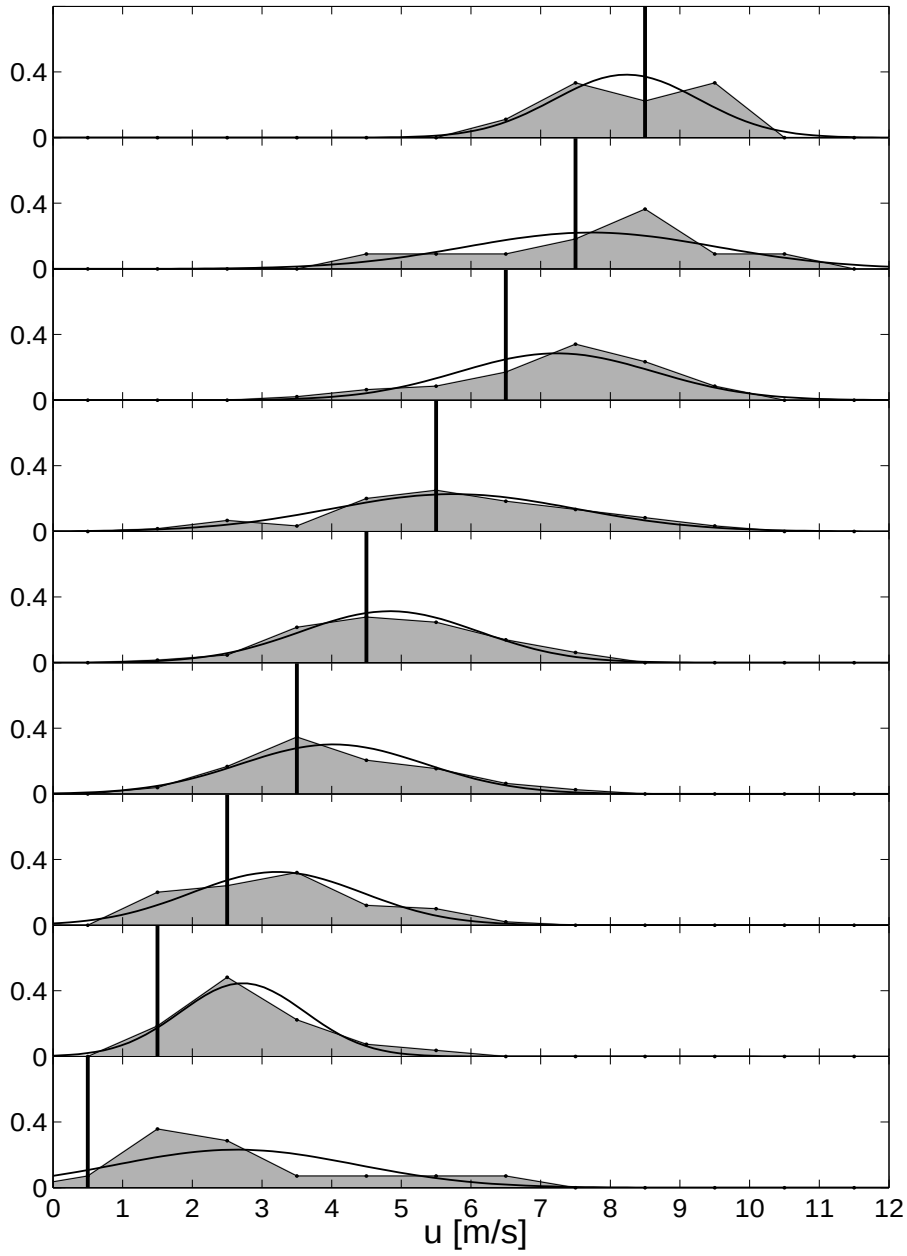

Figure 2.2: *Conditional probability density functions of 36 h wind speed values at site with low average wind speed (site 1). The different pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) (shaded areas) are stacked where $u_{\mathrm{pred,i}}$ has been varied in steps of 1 m/s, i.e. $\Delta u = 1$ m/s. The vertical line in each plot indicates the corresponding $u_{\mathrm{pred,i}}$. For comparison a normal distribution having the same mean and standard deviation as pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) is shown (solid lines). For pdfs with $2.5\ m/s \leq u_{\mathrm{pred,i}} \leq 6.5\ m/s$ the statistical tests indicate a Gaussian distribution. Note that conditional pdfs for large $u_{\mathrm{pred,i}}$ containing very few data points have been omitted.*

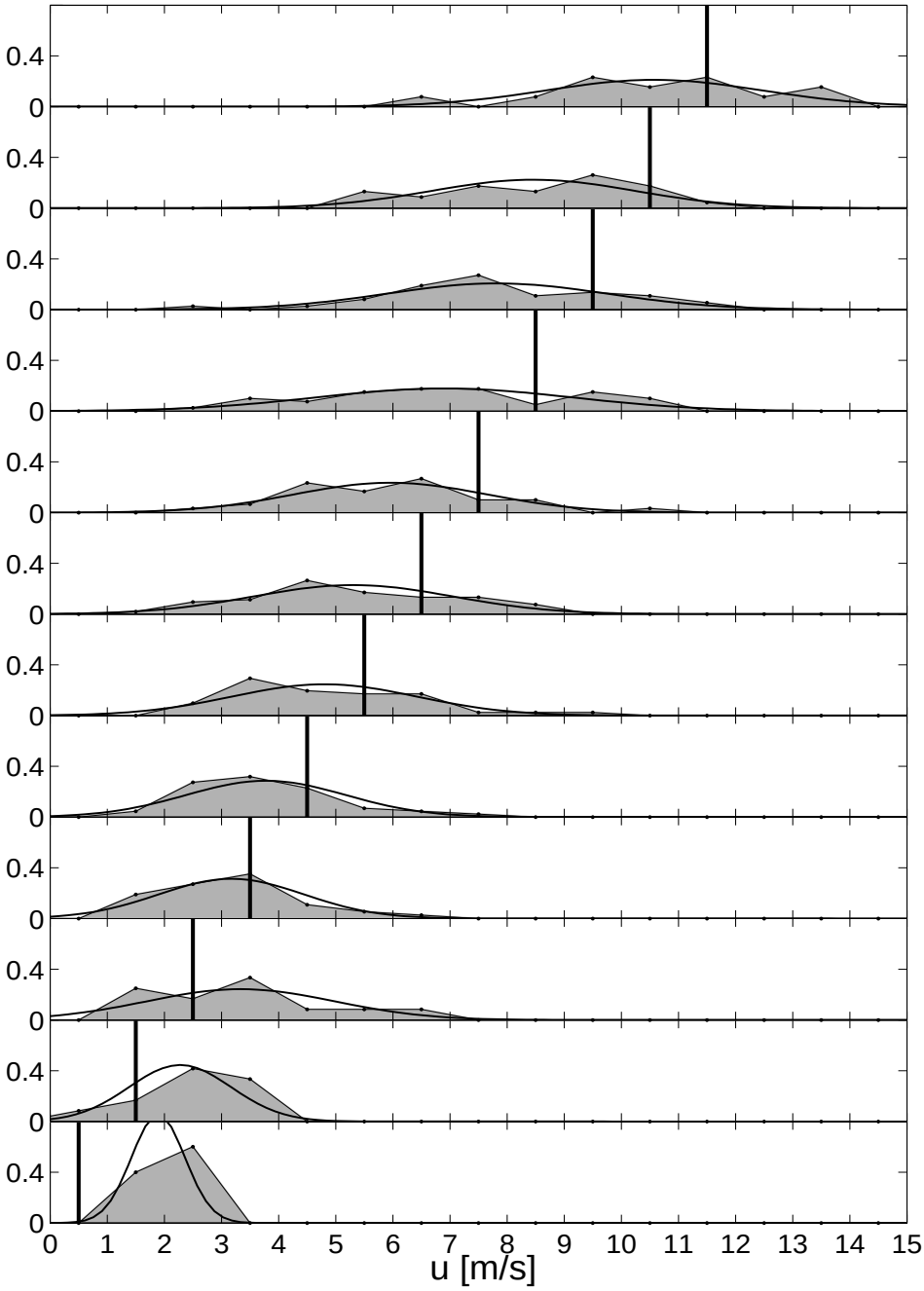

Figure 2.3: *As Figure 2.2 but for a high wind speed site (site 2). Again the pdfs tend to become similar to normal distributions with increasing $u_{\mathrm{pred,i}}$ which is confirmed by the tests for pdfs with $2.5\ m/s \leq u_{\mathrm{pred,i}} \leq 8.5\ m/s$. Note that the number of data points decreases for larger $u_{\mathrm{pred,i}}$ and conditional pdfs containing very few data points have been omitted.*

It is apparent from Figures 2.2 and 2.3 that in most cases the mean of the conditional pdfs does not correspond to $u_{\mathrm{pred,i}}$ which is related to the systematic errors in the data. This is further illustrated in Figure 2.4 where the means of the pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) are plotted versus $u_{\mathrm{pred,i}}$. For site 1 (Figure 2.4 (top)) the mean values are mostly above the diagonal, i.e. the predictions are on average smaller than the measurements, leading to a negative bias. In contrast to this the majority of mean values for site 2 (Figure 2.4 (bottom)) is below the diagonal which corresponds to an overall positive bias.

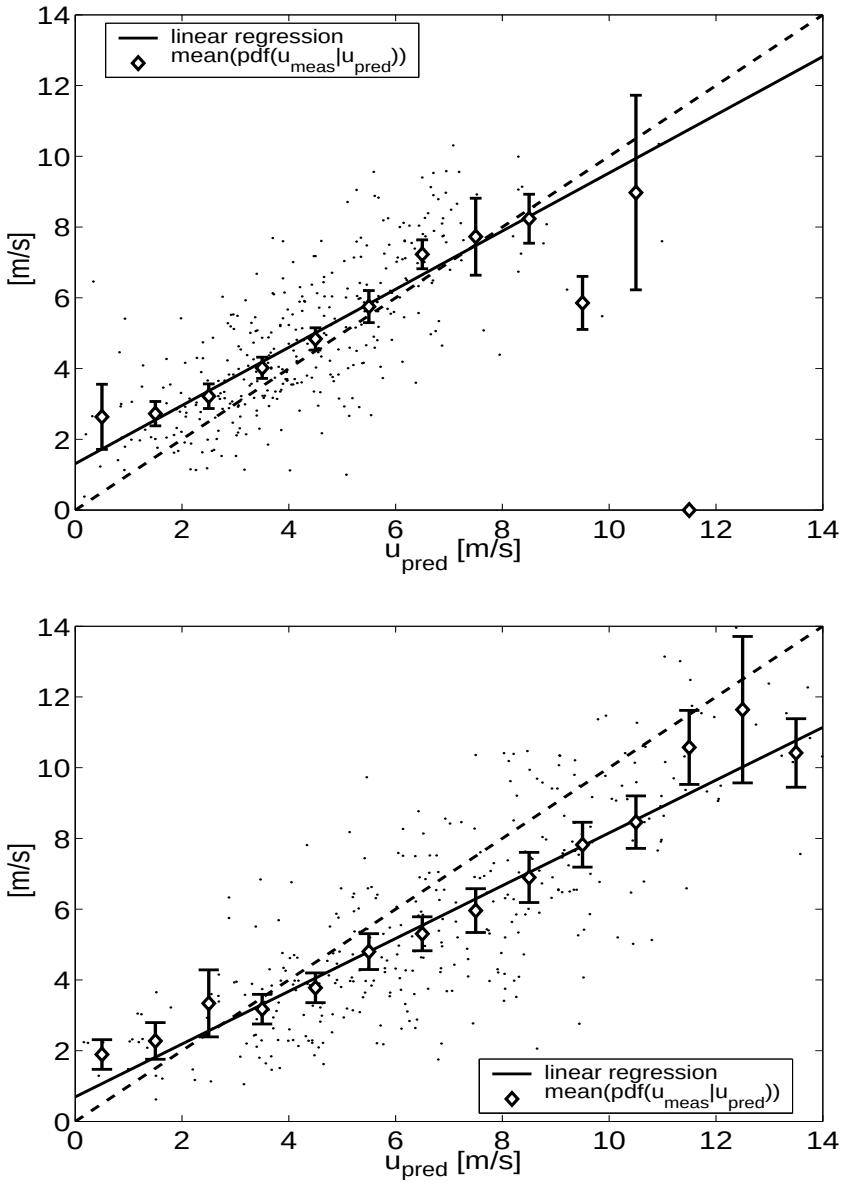

Figure 2.4: *The mean of pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) versus $u_{\mathrm{pred,i}}$ together with the actual data points $(u_{\mathrm{pred}}, u_{\mathrm{meas}})$ for the 36 h prediction. Top: site 1, bottom: site 2. In both cases the mean values are located on the linear regression line based on the data points for $u_{\mathrm{pred}} > 2\ m/s$. Their deviation from the diagonal (dashed line) reflects the systematic errors.*

For $u_{\mathrm{pred,i}} > 2\ m/s$ the mean values increase linearly with $u_{\mathrm{pred,i}}$ but with slopes different from unity indicating that the predictions systematically underestimate or overestimate the measurements. Within their error bars the mean values follow a line given by linear regression of all data points except for large $u_{\mathrm{pred,i}}$ where only few data points are available. Thus, the mean values of the conditional pdfs behave as expected in that they reflect, on a bin-wise level, the systematic errors of the complete time series.

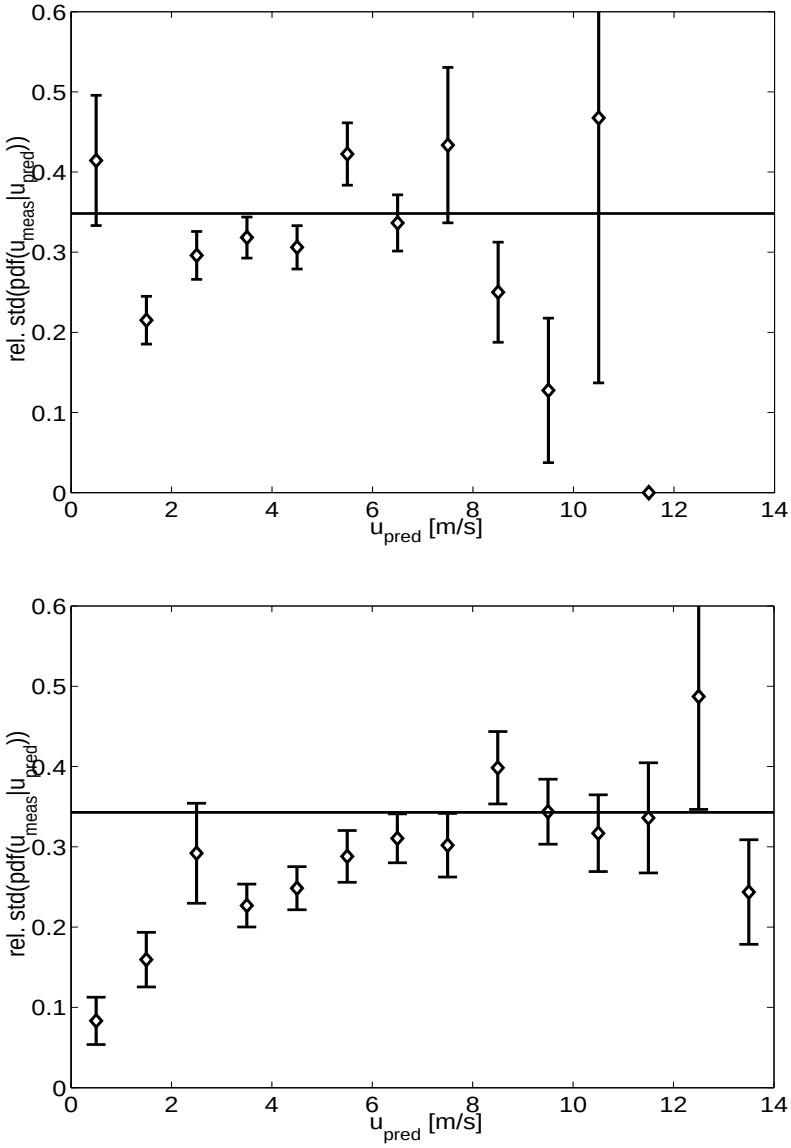

Figure 2.5: *The standard deviations of pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) normalised by the mean measured wind speed versus $u_{\mathrm{pred,i}}$ for site 1 (top) and site 2 (bottom). The prediction time is again 36 h. The solid line illustrates the unconditional relative standard deviation of the error. The error bars provide the confidence intervals in calculating the standard deviation of the conditional pdfs. While for site 1 no clear trend is detectable there seems to be an increase of the relative standard deviations of the conditional pdfs with the predicted wind speed for site 2. However, this is not systematic, neither for this prediction time at other sites nor for the other lead times at site 2.*

The pdfs in Figure 2.2 and 2.3 seem to become wider for increasing $u_{\mathrm{pred,i}}$ suggesting that the deviations of u_{meas} from $u_{\mathrm{pred,i}}$, i.e. the prediction errors, grow on average which is equivalent to a decreasing forecast accuracy. This directly leads to the question if the accuracy of the wind speed prediction depends on the magnitude of the wind speed. In this investigation no final answer can be given as the behaviour for different sites and prediction times is rather inconsistent.

Consider the two examples for site 1 (Figure 2.5 (top)) and site 2 (Figure 2.5 (bottom)). While at site 1 (top) the relative standard deviation of the pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) has no apparent trend and oscillates around the corresponding relative sde, the pdf of site 2 (bottom) seems to have an increasing standard deviation. But such a clear trend cannot be detected for other prediction times at this site.

Generally, most sites show some variation of the standard deviation of pdf($u_{\mathrm{meas}}|u_{\mathrm{pred,i}}$) over $u_{\mathrm{pred,i}}$ with relative standard deviations that deviate of the order 0.1 from the unconditional sde. The pdfs for $u_{\mathrm{pred,i}} \leq 2\ m/s$ are typically unsymmetric such that their standard deviation cannot be interpreted as 68%-confidence interval. Due to the limitation of these pdfs at the lower boundary their standard deviation is expected to be smaller compared to that one of the symmetric pdfs.

These considerations lead to the result that there is only a weak, if any, systematic dependence of the accuracy of the wind speed prediction on the wind speed.

2.4 Estimating the distribution of the power prediction error

In this section it is shown that the distribution of the power output can indeed be derived from the conditional pdfs of the wind speed together with the power curve. Based on the empirical pdfs of the wind speed found in the above section the unconditional distribution of the power prediction error can be reconstructed.

First of all, the quality of the reconstruction procedure is tested against a synthetic error distribution, denoted as $\mathrm{pdf}_{\mathrm{test}}(\epsilon_P)$, generated by using $\epsilon_P := P(u_{\mathrm{pred}}) - P(u_{\mathrm{meas}})$, i.e. using the theoretical power curve only. As Figure 2.6 illustrates for the case of site 1 the reconstruction of the error distribution that was calculated with Equations (2.9) and (2.11) and the conditional pdfs of the wind speed obtained in the last section (Figures 2.2 and 2.3) almost exactly recovers $\mathrm{pdf}_{\mathrm{test}}(\epsilon_P)$). Thus, though the procedure of first decomposing the wind speed data into conditional distributions, scaling them with the reciprocal derivative of the power curve and reassembling everything again leaves some space for numerical artefacts and inaccuracies due to small data sets it is robust enough to produce the expected results.

To be of practical use the more interesting test of the reconstruction is against the "real" distribution, $\mathrm{pdf}_{\mathrm{real}}(\epsilon_P)$, of the prediction error. Figure 2.7 shows the estimated $\mathrm{pdf}_{\mathrm{rec}}(\epsilon_P)$ for site 1 (top) compared to the distribution, $\mathrm{pdf}_{\mathrm{real}}(\epsilon_P)$, of the forecast error based on measurements of the actual power output. The overall agreement between the two distributions is rather good except for small ϵ_P. The reconstructed pdf covers the typical features of the original distribution in that it is unsymmetric in the same way and has the typical peak for small deviations.

As the error bars of $\mathrm{pdf}_{\mathrm{real}}(\epsilon_P)$ indicate the reconstructed distribution does not per-

fectly match the real one in all of the bins, in particular for site 2 (Figure 2.7 (bottom)). This deviation indicates additional sources of error that are not covered by considering the power curve effect only.

The results so far explain how the power prediction errors are statistically distributed and why the distribution has this special shape. If measured and predicted data of the power output are available there is no point in putting any effort in calculating the reconstructed $\mathrm{pdf}_{\mathrm{rec}}(\epsilon_P)$ as the error distribution is already available. However, if no data or only wind speed data is at hand the reconstruction can be used to get an idea how the error distribution of the power might look like. The minimum requirements to calculate this estimated distribution comprise four important aspects: First, the reasonable assumption that the conditional pdfs of the wind speed are all Gaussian (as discussed in section 2.3) with the same standard deviation, $\sigma(\epsilon_u)$. Second, a qualified guess concerning the value of $\sigma(\epsilon_u)$, e.g. from the weather service, as in [85], or nearby sites. Third, the distribution of the wind speeds at the desired site and, finally, the power curve of the wind turbine has to be available.

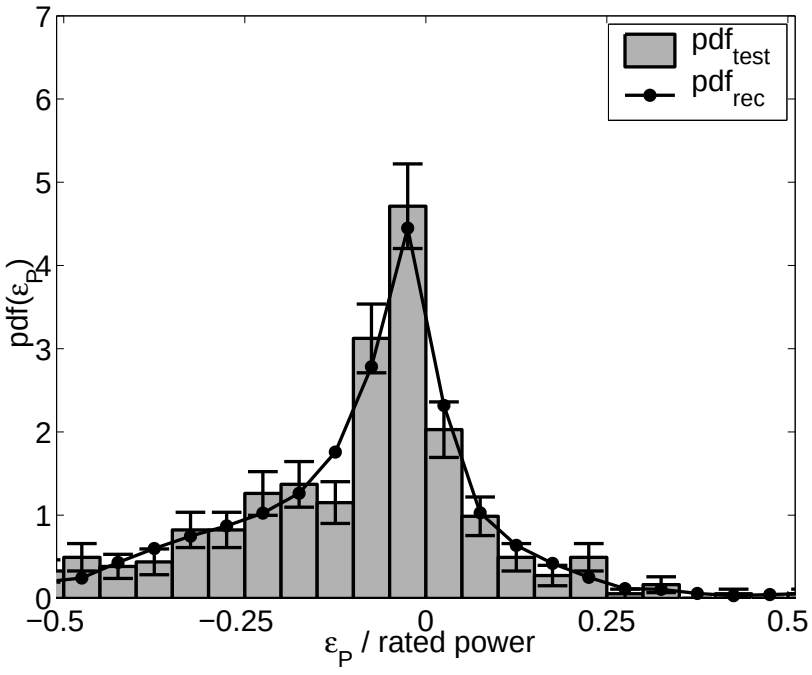

Figure 2.6: *Consistency check of the reconstruction procedure according to Equations (2.9) and (2.11) for the 36 h prediction at site 1. ϵ_P is normalised to the rated power. The reconstructed $\mathrm{pdf}_{\mathrm{rec}}(\epsilon_P)$ is compared to $\mathrm{pdf}_{\mathrm{test}}(\epsilon_P)$ which is based on evaluating $\epsilon_P := P(u_{\mathrm{pred}}) - P(u_{\mathrm{meas}})$. These two pdf should be identical. However, they show small deviations for the bins around $\epsilon_P = 0$.*

With the considerations of this section the conditional pdfs, $\mathrm{pdf}(\epsilon_P|P_{\mathrm{pred}})$, can in principle be constructed and, hence, for each prediction value P_{pred} an individual estimate of the error distribution around this value could be supplied. However, the data sets that are used do not allow for a proper verification of the individual $\mathrm{pdf}(\epsilon_P|P_{\mathrm{pred}})$ with measured data as, in particular, for medium and high power outputs the number of data points is

rather small. Hence, the difficulties in using statistical tests that already occurred for high wind speeds in section 2.3 are more severe with regard to power. Thus, in this investigation only the unconditional distribution of the power prediction error is reconstructed and compared to measurements because in this case all available data points at a site for a specific prediction time can be used.

In the next section it is shown that under simplifying assumptions a good estimation of the individual error bars of a specific wind power prediction can be derived.

2.5 Simple modelling of the power prediction error

Under the assumption that the prediction error of the underlying wind speed prediction does not change much over the range of typical wind speeds as discussed in section 2.3 the accuracy of the power forecast at a particular wind speed expressed as standard deviation of the error, sde, can be described rather well with a very simple approach.

For small deviations between predicted and measured wind speed the sde of the power prediction is linearised using a Taylor expansion at the predicted wind speed u_{pred} which leads to a "conditional" wind speed dependent error estimate given by

$$\sigma(\epsilon_P)|_{u_{\mathrm{pred}}} = \left|\frac{dP}{du}\right|_{u_{\mathrm{pred}}} \overline{\sigma(\epsilon_u)}. \tag{2.12}$$

This equation is rather basic in the sense that it takes only the effect of the local derivative $dP/du(u_{\mathrm{pred}})$ of the power curve into account and uses the average sde, $\overline{\sigma(\epsilon_u)}$, obtained from pdf$(u_{\mathrm{meas}}|u_{\mathrm{pred,i}})$ as accuracy for all wind speeds. The approach neglects that deviations between the certified, i.e. theoretical, power curve and the real power curve might occur which can have a significant influence on the prediction error. But this source of error can only be eliminated by using the corresponding measurement data to correct for these systematic deviations while Equation (2.12) generally models the error amplification effect of the power curve.

In Figure 2.8 the error of the power prediction where the corresponding predicted wind speed is confined to intervals of width 1 m/s for the two sites site 1 and site 2 is plotted versus the predicted wind speed. The bars denote the bin-wise standard deviation of the power prediction error, $\mathrm{sde}(\epsilon_P) = \sigma(\epsilon_P)|_{u_{\mathrm{pred}}}$ at a particular wind speed. Again one year of data has been used. Obviously, the accuracy of the power prediction depends on the predicted wind speed. This is mainly due to the power curve effect as the solid line calculated from Equation (2.12) indicates.

At site 1 this simple model describes rather precisely the behaviour of the actual power prediction error as illustrated in Figure 2.8 (top). However, at site 2 (Figure 2.8 (bottom)) this modelling approach does not lead to an accurate description of the prediction uncertainty. At this site the deviations between predicted and measured power output are not completely explained by the linear amplification of small wind speed errors according to Equation (2.12) due to differences between the certified and the real power curve. But for many sites the model provides a rather good estimation of the wind speed dependent power prediction error.

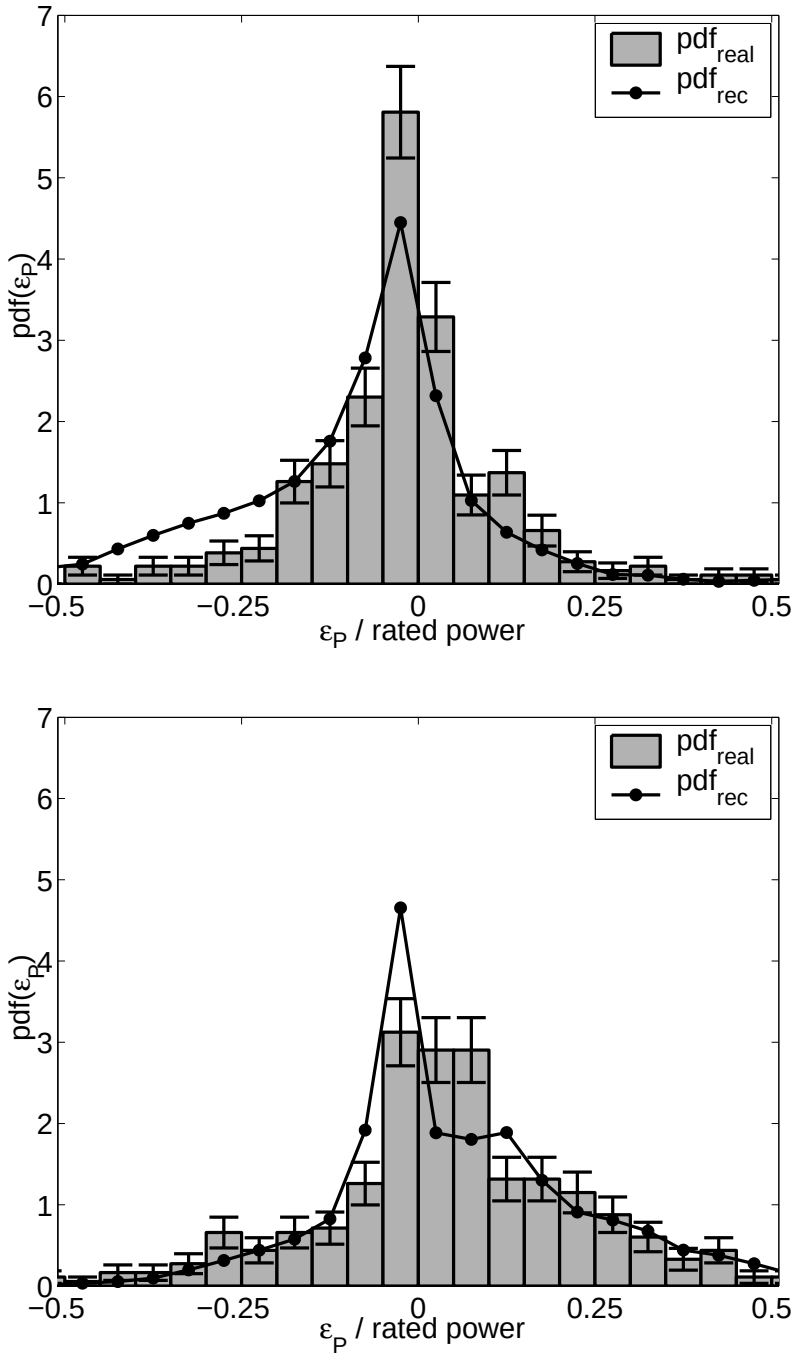

Figure 2.7: *Comparison of reconstructed* $\mathrm{pdf}_{\mathrm{rec}}(\epsilon_P)$ *according to Equation (2.11) to the distribution of the actual measured power output* $\mathrm{pdf}_{\mathrm{real}}(\epsilon_P)$ *as it was recorded for the 36 h prediction at site 1 (top) and site 2 (bottom).* ϵ_P *is normalised to the rated power. As the error bars of the real distribution indicate the agreement between the two pdfs for site 1 is rather good with* $\mathrm{pdf}_{\mathrm{rec}}$ *covering the typical features of* $\mathrm{pdf}_{\mathrm{real}}$ *except the large peak for very small* ϵ_P*. The two pdfs at site 2 show differences for small positive* ϵ_P *which indicate that the reconstruction model based on the power curve effect does not cover all error sources.*

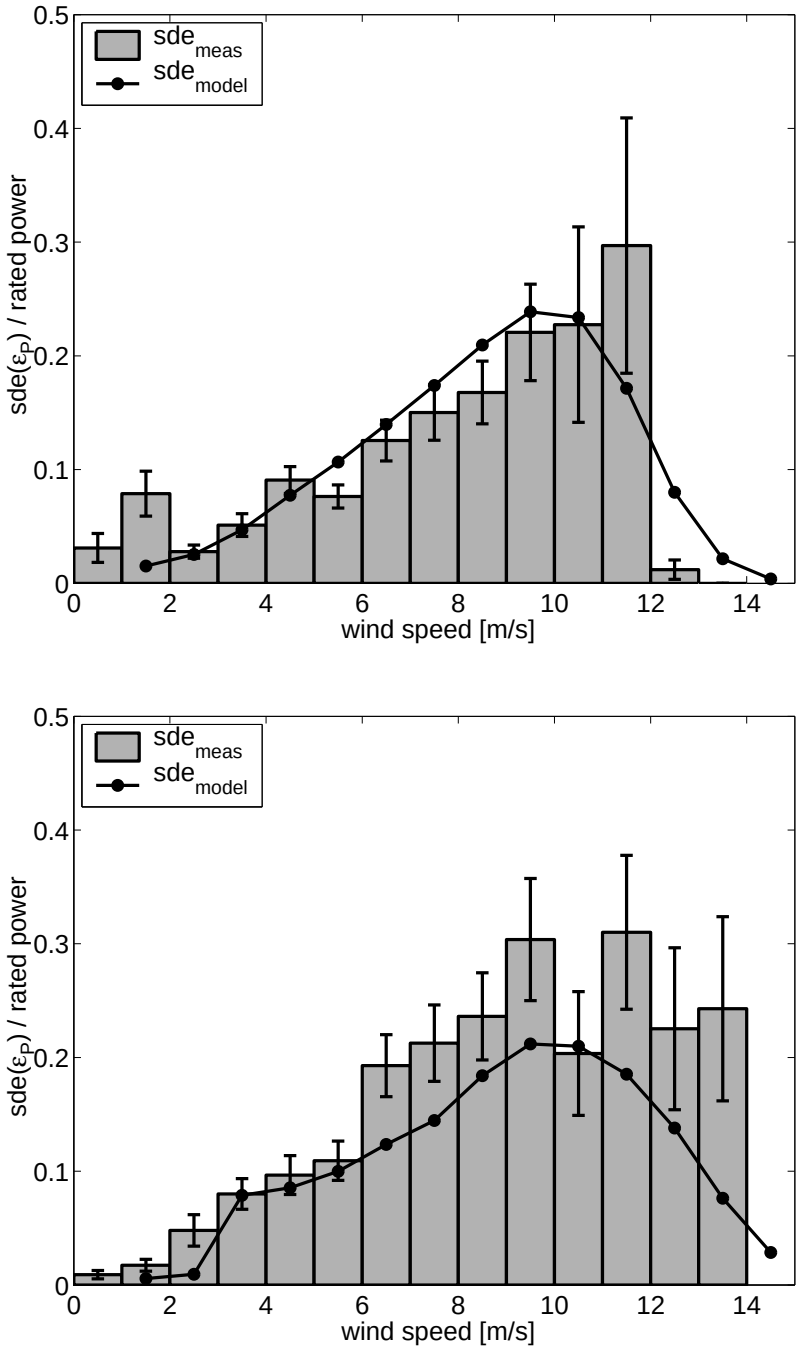

Figure 2.8: *Standard deviation of bin-wise power prediction error versus wind speed for the 36 h prediction at site 1 (top) and site 2 (bottom). The bars denote the sde between predicted and measured power output conditioned on wind speed intervals of width 1 m/s. For medium wind speeds up to 12 m/s the forecast error increases. At site 1 this behaviour is well approximated by the product of the derivative of the power curve and the wind speed error (solid line) according to Equation (2.12) while at site 2 the simple modelling approach underestimates the power prediction error for larger wind speeds which is due to effects not covered by this model.*

The amplification of the error caused by the local slope of the power curve is clearly detectable in the sde of the power forecast and can be used to model the prediction error depending on the wind speed. Figure 2.9 demonstrates how this can be implemented in a power prediction system like *Previento*. The specific uncertainty of each prediction given by Equation (2.12) is illustrated by the shaded area around the predicted value. It is calculated by using the derivative of the certified power curve and the $\overline{\sigma(\epsilon_u)}$ corresponding to each prediction time. Thus, to apply Equation (2.12) for prediction purposes historical data is needed to determine the statistical error, $\overline{\sigma(\epsilon_u)}$, of the underlying wind speed prediction.

This uncertainty interval provides additional information to the user and enables him to assess the risk of a wrong prediction. Of course, this procedure has to be refined in the future by taking the unsymmetric pdfs of the power error into account (see section 2.4).

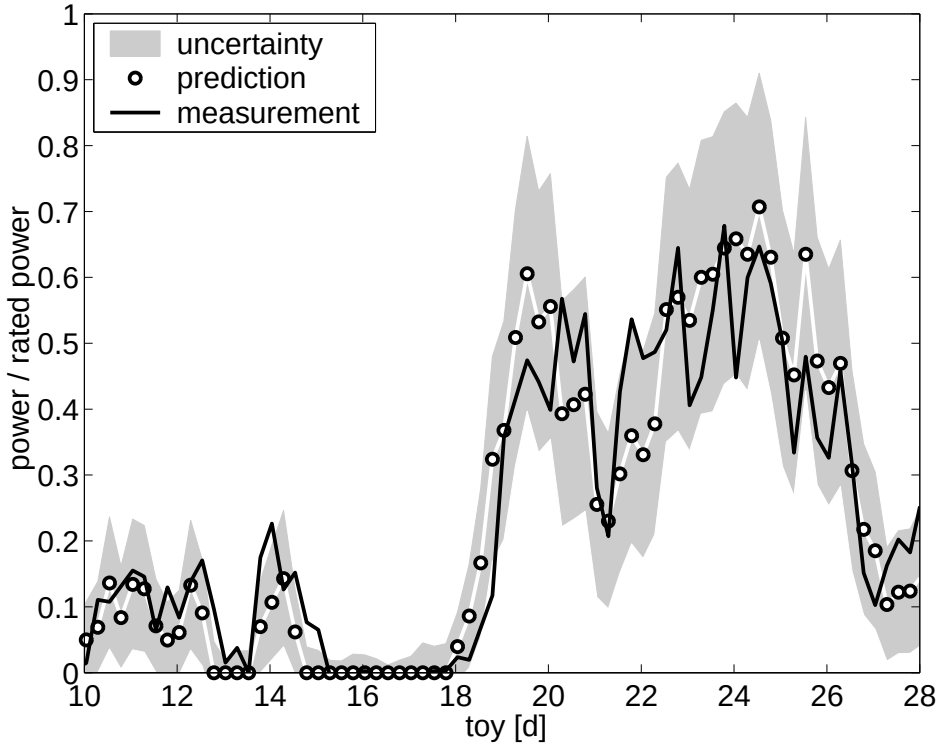

Figure 2.9: *Times series of prediction ("o") and measurement (solid line) over a period of 18 days. The shaded area is the uncertainty estimate given by Equation (2.12). Typically, the uncertainty interval around the predicted value is small for low predictions, i.e. flat slope of the power curve, and large for predictions between 20% to 80% of the rated power where the power curve is steep. Of course, for power outputs near the rated power (not shown here) the uncertainty again decreases.*

2.6 Conclusion

This chapter contains a first approach towards a situation dependent assessment of the prediction error where wind speed is the indicator that characterises the forecast situation.

Using wind speed as additional parameter a detailed analysis of the statistical properties of the prediction error shows that the conditional probability distribution functions (pdf) of the wind speed error are mainly Gaussian in the range of wind speeds that is important for wind power applications. The mean values of these pdfs are as expected on the line given by linear regression of the scatter plot reflecting the systematic error in the prediction. The more interesting question whether the prediction error expressed as standard deviation around these mean values increases with increasing wind speed cannot finally be answered here as the data for the different sites and prediction times does not show a consistent behaviour. However, the results of this chapter suggest that there is only a weak, if any, systematic dependence of the prediction error measured as the standard deviation of the differences between prediction and measurement on the magnitude of the wind speed.

A main result of this chapter is that the non Gaussian distribution of the power prediction error can be modelled. Understanding the mechanism that transforms initially Gaussian wind speed error distributions into strongly non Gaussian power error distributions allows to easily estimate the pdf of the power prediction error for any wind farm without knowing the actual predictions and measurements of the power output.

The approximated reconstruction of the pdf of the power forecast is based on three ingredients: the pdf of the measured wind speed conditioned on the predicted wind speed, the frequency distribution of the wind speed prediction and the power curve of the wind turbine. However, if for practical purposes the conditional pdf of the wind speed at a site is not available it can be estimated by assuming that the distribution is Gaussian with a standard deviation of the wind speed error taken from a nearby site or from the publications of the Weather Service (e.g. [85]). The exact frequency distribution of the predicted wind speed could be replaced by either the distribution of the measurements or distributions from nearby synoptic stations. This means that an estimate of the statistical distribution of the power prediction error at a site can be obtained using information that is readily available for existing wind farms and can even be obtained prior to the erection of the wind farm at the desired location. This is, for example, convenient to assess the prediction error for future offshore wind farms.

Of course, the "real" distributions of the power prediction errors given by comparing predicted and measured power output might look different. The method derived here does only consider the effect of the power curve neglecting other sources of error. In particular, the power output is modelled by plugging wind speeds into the certified power curve representing the average behaviour of the wind turbine under standard conditions. However, experience shows that the output of individual turbines can well deviate significantly from their certified power curves for various reasons [67]. Thus, for practical use the limitation of this reconstruction is clearly that it is oblivious to these additional error sources beyond wind speed. Nevertheless, it sheds some light on the mechanism that produces the typical non Gaussian distributions of the power prediction error.

A relatively simple model that describes the dependence of the power prediction error on the wind speed with a linearised approach is used. It shows that most of the error can be explained by the influence of the power curve that amplifies the rather constant prediction error of the wind speed according to its local derivative. This procedure can directly be implemented into *Previento* to provide situation dependent uncertainty estimates for each prediction time.

These uncertainty intervals are symmetric which is a major shortcoming of this simple model because it is already clear that the distributions of the power prediction error are non Gaussian and unsymmetric. However, to provide good estimations of the conditional power distributions around each prediction value a lot information has to be recorded concerning the statistics of the specific site. Therefore, the simple model suggested here is only the first step towards a comprehensive description of the situation based error statistics.

Chapter 3

Relating the forecast error to meteorological situations

Abstract

The investigation in this chapter focuses on the quantitative relation between the error of the wind speed prediction and the corresponding specific meteorological situation. With methods from synoptic climatology an automatic classification scheme is established using measurements of wind speed, wind direction and pressure at mean sea level to characterise the local weather conditions at a site. The classification procedure involves principal component analysis to efficiently reduce the data to the most relevant modes. Cluster analysis is used to group days with similar meteorological conditions into common classes. A comparison of these clusters with large-scale weather maps shows that typical weather situations are successfully captured by the classification scheme. The mean forecast error of the wind speed prediction of the German Weather Service is calculated for each of the clusters. It is found that different meteorological situations have indeed significant differences in the prediction error measured by a daily rmse where the maximum rmse can be by a factor of 1.5 to 1.7 larger than the minimum rmse. As expected, high uncertainties in the forecast are found in situations where rather dynamic low pressure systems with fronts cross the area of interest while stationary high pressure situations have significantly smaller prediction errors. This chapter is taken from [57].

3.1 Introduction

The investigation in this chapter continues to follow the idea of evaluating the forecast error for specific weather situations. But in contrast to the previous chapter the meteorological conditions will now be described by a far larger set of variables than only wind speed to include more details of the atmospheric state and its temporal evolution over one day. The aim is to really distinguish different weather classes and relate them quantitatively to their typical prediction errors.

It is a well-known fact that the performance of numerical weather prediction (NWP) systems is not equally well for every meteorological situation and that their accuracy depends on the situation that is to be forecast. In an overview of the prediction uncertainty of weather and climate forecasts Palmer [77] points out that “certain types of atmospheric flow are known to be rather stable and hence predictable, others to be unstable and un-

predictable". Thus, the challenge is to know in advance how predictable the current meteorological situation is.

There are already different approaches how to include information about the changing reliability of the numerical forecast into the prediction. A very popular one is the use of ensemble predictions where the chaotic properties of the non linear equations of motion of the meteorological variables are exploited, e.g. described in [77, 42]. Lorenz [60] demonstrated that low dimensional non linear weather models are sensitive to small changes in the initial conditions which is the typical indication of deterministic chaos. Hence, to get an overview over the possible range of weather situations that can evolve from a given situation the initial condition of the NWP is perturbed. Then the NWP calculates separate predictions for each initial condition leading to an ensemble of possible outcomes. Thus, ensemble forecasts provide a range of possible weather situations that can occur with a certain probability. The difficult part is to generate suitable ensembles allowing for a statistical interpretation of the results which might differ profoundly.

Ensemble forecasts are computationally expensive and, therefore, for operational use mainly generated by large meteorological institutes such as the European Centre for Medium Range Weather Forecasts (ECMWF). However, Landberg et al. [54] used a simpler version of ensemble predictions in connection with wind power application which they denoted as "poor man's ensemble forecast". The idea is to take the spread between different prediction runs, e.g. at 00 UTC and 12 UTC, calculated for the same points of time in the future. The larger the deviation between the prediction runs the greater the uncertainty of the prediction. This concept has recently been enhanced by Pinson and Kariniotakis [80, 79]. They established a continuous risk index from the spread of several forecast runs of the wind speed prediction to derive the corresponding degree of uncertainty connected to the power prediction.

The technique used in this chapter to assess the uncertainty of a specific prediction is different from the ensemble approach as it is based on a classification scheme of the weather situation which is not directly related to the prediction system. The approach here is to describe the situation of each day by a suitable set of local meteorological variables, then sort it into a certain category, and associate each of the categories with a typical prediction error derived from historical data.

Intuitively, at least two types of meteorological situations over Northern Europe are expected to show considerable differences in terms of the accuracy of the wind speed prediction. Weather situations with strong low pressure systems coming in from the Atlantic Ocean are supposed to be of the unstable type that is hard to predict. These situations can be very dynamic as the advection speed of the low is typically rather large and, in addition, its frontal zones may cause strong winds. In dynamic cases the real situation can evolve quite differently from the one that had been predicted. In contrast to this high pressure systems with typically moderate wind speeds are rather stationary and can persist for several days which should make it easier for the NWP to provide a reliable prediction.

It is the purpose of this chapter to establish a method that automatically generates useful weather classes and, hence, implicitly defines what "strong low" or "stationary high" means in terms of the local conditions at different sites. Moreover, the investigation

here aims to answer the question how large the various prediction errors are and whether they differ significantly for different meteorological conditions.

3.2 Methods from synoptic climatology

The techniques that are applied in this investigation are well established in synoptic climatology. In a concise overview over the subject Yarnal [94] points out that "synoptic climatology relates the atmospheric circulation to the surface environment". Hence, this branch of climatology explicitly aims at linking meteorological conditions to external variables, such as concentration of air pollutants, which is exactly the type of problem that is addressed in this work.

Quite a variety of powerful methods has been developed in this field and tested in numerous applications [4, 94, 86]. The main principles are very similar. The first step is to identify typical structures or patterns of the atmospheric circulation in order to develop a classification scheme that is only based on meteorological variables. After the classification is established the statistical or deterministic relationship between these structures and a surface variable of interest is investigated. Normally, this variable is non meteorological in the sense that it does not describe the state of the atmospheric flow.

Regarding the classification scheme there are two main approaches. On the one hand manual methods where a meteorologist evaluates the synoptic situation within a certain framework according to his experience and on the other hand automatic schemes which use computer-based algorithms to sort meteorological data into different classes. A well-known manual classification scheme for Central Europe is the catalogue of "Grosswetterlagen" by Hess and Brezsowski [31]. It comprises 29 large scale weather situations which are distinguished by their different spatial arrangement of the pressure systems. One disadvantage in using this cataloque is that the large number of pre-defined weather classes requires times series with many data points in order to obtain statistically relevant results [58].

In contrast to this, automatic schemes typically exploit correlations between patterns or use eigenvector techniques such as principal component analysis (PCA) to extract synoptic weather classes from numerical data. Yarnal [94] points out that both the manual and the automatic methods contain a certain degree of human subjectivity because someone has to decide, e.g, on the set of variables and the number of classes.

The method chosen in this work to analyse how the forecast error is related to the weather type relies on computer-based classification techniques and follows in principle an investigation carried out by Shahgedanova et al.[86]. They used principal component analysis in combination with cluster analysis of surface and upper air meteorological data to derive a classification scheme of the synoptic situation in Moscow where PCA was used for data reduction and cluster analysis to sort similar days into common groups. These weather types were then connected to typical concentrations of air pollutants such as CO and NO_2. It was found that certain weather patterns result in significantly high levels of urban air pollution.

In contrast to Shahgedanova et al. the set of meteorological variables used here will be smaller with no upper air data involved. The role of the pollutant will be played by

the forecast error. In the following the use of PCA and cluster analysis is explained.

3.2.1 Principal component analysis (PCA)

Principal component analysis (PCA) is used to extract the relevant eigenmodes from the meteorological data. This technique produces eigenmodes which can be ordered according to the degree of variance they explain in the data. Hence, the modes that contribute most to the time series can be selected for the further analysis which allows to effectively reduce the amount of data. The weather situation of each day can then be approximately expressed as a linear combination of these eigenmodes.

The variables used here are the horizontal wind vector $\vec{u} = (u, v)$ and the atmospheric pressure at mean sea level (pmsl). These are the natural candidates to start with as they are closely related to the wind field. To account for temporal variations over one day the variables are taken at 0, 6, 12, 18, 24 UTC. Several values of the wind vector per day resolve changes in wind direction and speed, e.g. during the passage of a frontal system. This provides the possibility of separating dynamic from static weather situations. The record of the temporal variations of pmsl can partly compensate for the fact that spatial pressure gradients which are the main driving force of the wind are not available in this investigation.

The temperature that is often included to find synoptic indices is not considered here. This is done because absolute temperature is mainly used to determine the type of air mass, e.g. by Shahgedanova et al. [86], but it is not directly related to the wind field. Moreover, temperature tends to be a dominating variable that requires the data sets to be split into winter and summer part. But doing so would leave only half the data points for the analysis which makes it difficult to have significant results in the end. However, for further investigations it would be desirable to include temperature and humidity to be able to detect fronts with more parameters than changes in the wind vector. Of particular interest for further investigations are, of course, temperature differences at different heights to assess atmospheric stratification (c.f. Focken [29]).

The measured data are written into a matrix M where the columns contain the different variables and each row corresponds to one day. The measurements at the various day times are considered as separate variables. As wind vector and surface pressure have very different orders of magnitude both are normalised with their standard deviation and pmsl is additionally centralised by subtracting the mean value. Hence, M is given by:

$$M = \begin{pmatrix} u_{1,0} & \cdots & u_{1,24} & v_{1,0} & \cdots & v_{1,24} & \mathrm{pmsl}_{1,0} & \cdots & \mathrm{pmsl}_{1,24} \\ \vdots & & & & & & & & \vdots \\ u_{365,0} & \cdots & u_{365,24} & v_{365,0} & \cdots & v_{365,24} & \mathrm{pmsl}_{365,0} & \cdots & \mathrm{pmsl}_{365,24} \end{pmatrix} \tag{3.1}$$

where the subscripts denote the day of the year and the day time. M is the so-called data matrix with dimension 365 by 15 having one row for each day of the year with 15 measurements per day at the times 0 h, 6 h, 12 h, 18 h and 24 h UTC.

The PCA of matrix M is carried out numerically by diagonalising the so-called covariance matrix $C = M^t M$. The 15 eigenvectors, $\vec{p}_i$, of C are the desired principal com-

ponents (PC) of M while the eigenvalues, λ_i, of C are the weights λ_i of the different components that express the degree of variance a particular PC contains. A full description of this procedure can be found in Broomhead and King [11].

The set of principal components $\{\vec{p}_i\}$ with $i = 1, \ldots, 15$ constitutes a new orthonormal basis of the phase space of the full set of data points. The ordering of the PC according to the eigenvalues, λ_i, allows to select only the first few PC spanning the subspace where most of the relevant dynamics takes place. The number of relevant modes, N, to be considered for further analysis is not precisely defined and has to be inferred from the spectrum of eigenvalues and the corresponding PC. In this investigation the visual inspection of these quantities and the cumulative variance $\sum \lambda_i$ of the first N eigenvalues will be used as criteria to decide on the number of modes to be used. Hence, by transforming to new coordinates and reducing the degrees of freedom PCA efficiently codes the information contained in the meteorological raw-data and in that sense acts as a data reduction technique.

The use of PCA as a data reduction technique means that the eigenvalue spectrum is mainly used to account for the degree of variance contained in each of the corresponding eigenmodes. This is a statistical or climatological interpretation of the eigenspectrum and not a non linear systems approach. In the context of dynamic systems a similar technique, time-delay embedding, is also applied as a tool to decompose phase space dynamics in different modes with the aim of extracting the degrees of freedom of the non linear system, e.g. described in detail by Broomhead and King [11] or Kantz and Schreiber [47]. However, the approach used in this work and in particular the construction of the data matrix M does not aim at providing a delay embedding of the time series as the sampling intervals are not adapted for this purpose and the number of data points is far too low. Hence, the number of eigenmodes provided in this context cannot contribute to the question how many degrees of freedom the weather or climate system has as discussed, e.g, by Nicholis and Nicholis [69] and Grassberger [35].

After a choice concerning the number of relevant PC is made, the data in M is transformed to the reduced basis such that each day can be represented as a linear combination of the selected PC. Let $\{\vec{q}_i\}$, $i = 1, \ldots, N$, be the basis chosen from the first N PC of the full eigenvector basis, $\{\vec{p}_i\}$, of M. Then

$$Q := (\vec{q}_1 \ldots \vec{q}_N) \tag{3.2}$$

is the transformation matrix that can be used to easily project the data in M on the new basis by a multiplication from the right

$$X := MQ. \tag{3.3}$$

The entries in X are the scalar products $x_{ij} = \vec{m}_i \cdot \vec{q}_j$ where $\vec{m}_i$ is the i-th row of M. In other words x_{ij} is the contribution of the j-th PC to the i-th day. Consequently, each day $\vec{m}_i$ can be approximately (because only a N-dimensional subspace is considered) expressed by

$$\vec{m}_i \approx \sum_{j=1}^{N} x_{ij} \vec{q}_j. \tag{3.4}$$

Thus, the 365 by N matrix X is the reduced data matrix that is thought to contain most of the relevant meteorological information of one year of data. However, so far nothing really happened in terms of the classification scheme because the data has merely been recoded. The next step applied to the reduced data set will be cluster analysis.

3.2.2 Cluster analysis

Cluster analysis is a standard method used to group objects with similar properties together. As described in a concise overview by Everitt [27] it has a wide range of techniques and is often applied in climatological investigations [94, 46, 86].

In this work the aim is to obtain a rather small number of clusters which contain days with similar meteorological conditions being different from the days in the other clusters. Of course, the clusters found are required to represent typical weather classes.

The type of cluster analysis used here is called hierarchical or agglomerative. It acts iteratively on the phase space by computing the distances between each pair of objects in the phase space and then joining the closest two objects to a new cluster. After the new cluster is formed the procedure is repeated. Starting point is a situation where all points in the phase space are considered as separate clusters. As the final iteration merges all clusters into one group the process has to be stopped when a certain number of clusters is reached. The criterion to stop the iterations is derived from observing the growing distances between the clusters.

Though the basic concept is rather straightforward there are profound differences in the results of various clustering procedures because the key point of this technique is how distances between clusters are defined and under which rule new clusters are formed. First of all, a metric has to be chosen that evaluates distances between single points in the phase space. Here the Euclidean metric (Equation (3.5)) is applied as there is no apparent reason for a different one. A more crucial point is the definition of distances between clusters. As the type of distance measure implies which two clusters will be joined next it is referred to as "linkage method". Three typical linkage methods are briefly introduced using the following notations.

Let $\vec{x}_i := (x_{i,1}, \ldots, x_{i,N})$ with $i = 1, \ldots, 365$ denote the coordinate vector describing a point in the N dimensional reduced phase space. The distance $d(\vec{x}_i, \vec{x}_j)$ between two states is given by the Euclidean measure

$$d(\vec{x}_i, \vec{x}_j) = \sqrt{\frac{1}{N} \sum_{q=1}^{N} (x_{i,q} - x_{j,q})^2}. \tag{3.5}$$

After a number of N_A phase states has been joined into one cluster, C_A, the members of this cluster will be denoted as $\vec{x}_r^A$ where $r = 1, \ldots, N_A$ is the new index within C_A.

Complete linkage method

To evaluate the distance between two clusters C_A and C_B in the complete linkage method all pairs of Euclidian distances between members of C_A and C_B are computed. Then the maximum distance that is found between the individual members is used as distance

between the two clusters. Hence,

$$d_{\text{complete}}(C_A, C_B) := \max\left(d(\vec{x}_r^A, \vec{x}_p^B)\right) \quad \text{where} \quad (\vec{x}_r^A, \vec{x}_p^B) \in C_A \times C_B. \tag{3.6}$$

Average linkage method

Average linkage is a combination of complete linkage as described above and its complementary definition, called single linkage, where the minimum of the distance between cluster members is taken. In average linkage the mean value of the individual distances is used to define the cluster distance between two clusters C_A and C_B, i.e.

$$d_{\text{average}}(C_A, C_B) := \frac{1}{N_A N_B} \sum_{r=1}^{N_A} \sum_{p=1}^{N_B} d(\vec{x}_r^A, \vec{x}_p^B) \tag{3.7}$$

where N_A and N_B are the respective numbers of elements in C_A and C_B.

Ward's linkage method

Ward's method is also referred to as minimum variance method [46] as it evaluates the change of within-cluster variance if two clusters are merged.

$$d^2_{\text{ward}}(C_A, C_B) := \frac{N_A\, N_B}{N_A + N_B}\; d\left(\overline{\vec{x}^A}, \overline{\vec{x}^B}\right)^2 \tag{3.8}$$

where $\overline{\vec{x}^A} = \sum_{r=1}^{N_A} \vec{x}_r^A$ is the centre of mass of cluster C_A and $\overline{\vec{x}^B}$ the one of C_B.

Kalkstein et al. [46] thoroughly investigated average linkage, Ward's linkage and a third technique, centroid linkage, that is not used here. They found that for weather classification purposes average linkage produced the "most realistic synoptic groupings" as it provides rather homogeneous groups of days with similar meteorological conditions and sorts extreme events into separate clusters instead of combining them into a common class. In contrast to the other techniques average linkage minimises the within-cluster variance, i.e. the mean variance among the days within one cluster, and maximises the between-cluster variance, i.e. the mean variance between the centres of mass of different clusters. Shahgedanova et al. [86] also confirmed the usefulness of average linkage in their investigation.

Ward's linkage on the other hand tends to produce clusters of equal size and, therefore, sorts days with extreme weather conditions together with less extreme days which "blurs distinctions between the types" [46]. Thus, this method was considered as inferior compared to average linkage to produce meaningful synoptic classes. However, Yarnal [94] provides classification problems which are quite comparable to those by Kalkstein et al. [46] and Shahgedanova et al. [86] but where Ward's method is superior to average linkage which leads the conclusion that both methods should be tested.

In this work average and Ward's linkage are used as they have been successfully applied in previous investigations. Additionally, complete linkage is also applied.

3.2.3 Daily forecast error of wind speed

For the analysis of the forecast error a new error measure is introduced that evaluates the performance of the prediction system over one day:

$$\mathrm{rmse}_{\mathrm{day}} = \sqrt{\frac{1}{4}\sum_{t_{\mathrm{pred}}} \left(u_{\mathrm{pred},t_{\mathrm{pred}}} - u_{\mathrm{meas},t_{\mathrm{pred}}}\right)^2} \quad \text{with} \quad t_{\mathrm{pred}} = 6\mathrm{h}, 12\mathrm{h}, 18\mathrm{h}, 24\mathrm{h} \tag{3.9}$$

where $u_{\mathrm{pred},t}$ and $u_{\mathrm{meas},t}$ are predicted and measured wind speed.

It is, of course, desirable to also include predictions with horizons beyond 24 h into the error measure. However, higher lead times have not been included so far because only two larger prediction times, 36 h and 48 h, are available in the investigated data set which is only half the number of data points for the analysis compared to the period 6 to 24 h. Hence, this first approach is restricted to the earlier lead times.

This "daily error" assigns one single value to each day in a cluster which is analogous to using the daily concentration of a pollutant by Shahgedanova et al. [86]. The characteristic property of this definition is that the point of time at which a deviation between prediction and measurement occurs is not relevant. Moreover, with the definition in Equation (3.9) errors caused by coherent structures such as a wrongly predicted front are summarised in one error value, i.e. the correlation between succeeding deviations is implicitly taken into account.

The idea behind choosing $\mathrm{rmse}_{\mathrm{day}}$ can be illustrated by an example. Imagine a low pressure system approaches the domain of interest. It brings a rise in wind speeds that is predicted for the late afternoon, say 18 h, but actually arrives a few hours earlier, say 12 h. In this case the difference between predicted and measured values is rather large and negative at 12 h (because the prediction did not foresee the low) and small and negative at 18 h (because the prediction expected the wind speed to start increasing). Moreover, at 24 h the low has passed and the wind speed actually decreases while the prediction still suggests high wind speeds leading to a positive deviation. Now these situations can typically produce phase errors where the deviations at a number of lead times in a row are coherently affected. Another important point is that in terms of the daily error measure it should not matter whether the low is too fast as in this example or is behind schedule, i.e. arrives later than predicted.

3.2.4 Tests of statistical significance

Resolving individual situations reduces the amount of data points available per cluster by about a factor 10 and, thus, statistical significance becomes an important issue. Therefore, care has to be taken not to be fooled by statistical artefacts such that in this chapter and the next one tests for statistical relevance will again be applied. To check if the differences in the statistical values of different clusters are not obtained by chance an F-test together with Scheffe's multiple comparisons are applied which are based on relating the within-cluster variance of the error values to the variance of the error values in the remaining clusters.

3.3 Results

The methods described above are applied to measurement data at different sites. To develop a climatological classification scheme measured data from one year is used. The horizontal wind vector $\vec{u}$ at a site is calculated by sine-cosine transformation of the measured wind speed and direction at, preferably, 30 m to avoid artefacts from obstacles that might occur for lower heights. However, if the 30 m measurement is not available 10 m data are also used. The corresponding surface pressure is taken from the nearest synoptic station of the German Weather Service. It is converted to mean sea level using the barometric height formula.

To avoid complications due to missing data points only sites with a high data availability in wind speed, wind direction and surface pressure are chosen. The investigation here will focus on two locations Fehmarn and Hilkenbrook with 30 m wind data. Fehmarn is an island in the Baltic Sea close to the German coast. The site that is investigated is near the island's shore line and, hence, exposed to rather inhomogeneous conditions. Northerly and north westerly winds approach the site over the sea while easterly to south westerly winds arrive over land. In contrast to this Hilkenbrook is in the north western part of Germany about 70 km south of the coast of the North Sea. The terrain is rather homogenous in terms of the surface roughness and no orographic effects are to be expected.

The results are presented in the following way. First the modes extracted from the meteorological variables with PCA are presented. Then cluster analysis is applied to the reduced data matrix and the typical weather classes that are obtained by a straightforward approach are shown for complete as well as Ward's linkage and compared to large-scale weather maps. The average forecast error of the wind speed for each cluster is provided and discussed for the different linkage techniques.

3.3.1 Extraction of climatological modes

The PCA as described in section 3.2.1 is carried out numerically using routines from the standard software MATLAB. In Figure 3.1 the eigenvalues, λ_i, of the PCA for Fehmarn, Hilkenbrook and, for comparison, another site, Rapshagen, are shown. The eigenvalues are normalised with the total variance, i.e. the sum of all eigenvalues. It can be clearly seen in Figure 3.1 that the eigenvalues decay rapidly with the first six eigenvalues already explaining about 90 % of the variance.

The spectra are surprisingly similar though the stations are in regions with distinct weather regimes and different local conditions around the site. There are only minor deviations for the first and most important eigenvalues, i.e. the distribution of variance among the corresponding PC is rather consistent. The eigenvalues for Syke are not shown as they are as expected almost identical to those of the rather adjacent site Hilkenbrook.

The first six principal components (PC) in $\vec{u}$-space of Fehmarn and Hilkenbrook are shown in Figure 3.2. The temporal evolution of $\vec{u}(t) = (u_t, v_t)$ at times $t = 0, 6, 12, 18 \text{ and } 24$ h is indicated by connecting the end points (u_t, v_t) with lines. Note that the wind vector, i.e. the direction in which the wind blows, is shown while the description of wind directions will refer to the direction from which the wind comes. Hence, e.g. southerly winds refer

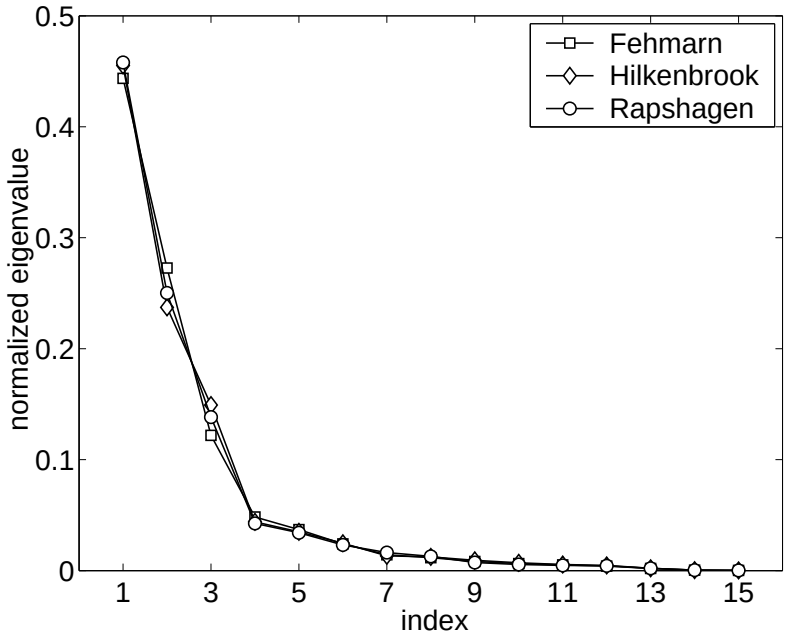

Figure 3.1: *PCA eigenvalue spectrum normalised with the sum of the eigenvalues, i.e. the total variance, for three sites (Fehmarn, Hilkenbrook and Rapshagen). Though the sites have different local wind characteristics the spectra are similar.*

to wind vectors pointing to the north.

The corresponding PC of pmsl are illustrated in Figure 3.3. Note that the PC do not necessarily correspond to actual meteorological conditions at the location on a specific day. They constitute the set of vectors whose linear combination can approximately reconstruct the current situation (cf. Equation (3.4)).

Typically, the first two PC are rather stationary in that they show little diurnal variation in terms of wind direction and pressure. For all investigated sites the first PC describes moderate wind speeds from east or north east with slightly higher wind speeds at noon and virtually no change in wind direction over the day (see Figures 3.2 and 3.3). In addition, the pmsl is at a constantly high level. PC 2 is also quite consistent for different locations. It refers to north westerly wind directions with increasing wind speeds at midday. The corresponding pmsl is slightly rising at a high level. In terms of $\vec{u}$ these first two PC can roughly be associated with the two most frequent flow directions at the site.

The third PC is again similar for all sites although it occurs with different signs. As the sign of the PC is arbitrary, those PC which only differ in their sign are regarded as identical climatological modes. In half of the cases PC 3 refers to wind from the north with pmsl increasing from low to medium level while for the other half it describes wind from the south in connection with pmsl falling from high to medium level.

Higher principal components are much more dynamic for all sites than the first three PC. As can be seen in Figures 3.2 and 3.3 the diurnal variations in all variables are profound. PC 4 to 6 typically refer to changes in wind direction of about 180 deg up to

300 deg with rapidly changing wind speeds and pmsl. Hence, these PC are needed to account for changing meteorological conditions over one day.

For adjacent stations the PC are almost identical. However, even for sites that are supposed to be located in different climatological conditions like Fehmarn and Hilkenbrook the sets of relevant PC are surprisingly similar. This suggests that PCA extracts fundamental modes of the climatology that are quite universal for the investigated area of northern Germany.

The inspection of the eigenvalue spectra in Figure 3.1 reveals that the first four PC on average contribute most to the meteorological signal. However, the structure of the PC as discussed above showed that, in particular, the PC 4 to 6 describe the dynamic changes of the weather condition occurring within 24 hours. Hence, these higher modes are expected to represent more extreme meteorological situations which do not occur often but are supposed to account for larger forecast errors than more stationary modes. Consequently, the first six PC are chosen, i.e. $N = 6$, as the orthogonal basis $\{\vec{q}_i\}$, $i = 1, \ldots, 6$, spanning the reduced state space.

Hence, the number of modes chosen here is slightly larger compared to what the eigenspectrum suggests. Yarnal [94] points out that the use of too many eigenmodes does not necessarily enhance the performance of the following cluster analysis which is quite understandable as the additional variation introduced by including more modes might not contain useful information but mainly noise. However, in the course of this investigation it will turn out that in particular the PC which describe changes in the meteorological variables over one day are important. The use of less than six PC has so far not been investigated.

According to Equations (3.2) and (3.3) the data matrix M is projected onto the new basis and the day-score matrix X is obtained that contains the contributions of the six PC to the meteorological situation of a specific day.

3.3.2 Meteorological situations and their forecast error

The 365 by 6 day-score matrix X is the basic element of the cluster analysis as its rows constitute the phase space points that are to be grouped together. The three linkage techniques average, complete and Ward's linkage have been used in this investigation to define clusters. In contrast to Shagedanova et al. [86] and Kalkstein et al. [46] average linkage as defined in Equation (3.7) failed in producing meaningful synoptic classes as it sorted two thirds of the days into one common group and divided the remaining days into a number of small groups. Complete linkage (Equation (3.6)) and Ward's linkage (Equation (3.8)) produce clusters that can be associated with typical weather situations and, hence, the classification schemes based on these linkage types appear to be reasonable. In the following the results of complete and Ward's linkage are presented for the two sites Fehmarn and Hilkenbrook.

Complete linkage

The first step of the cluster analysis is to successively join the objects in the phase space starting from 365 separate days up to one single cluster and record the distances accord-

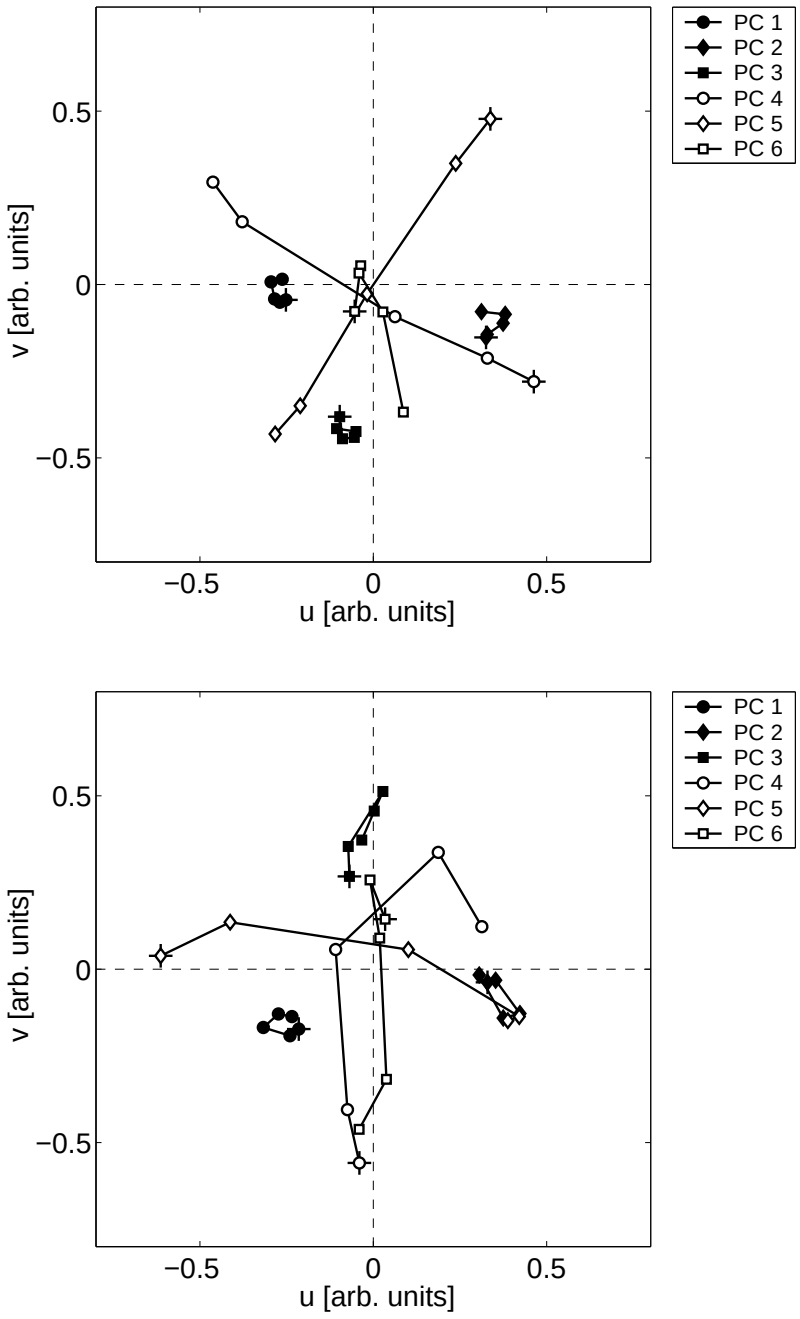

Figure 3.2: *First six principal components (PC) of $\vec{u}$ space for Fehmarn (top) and Hilkenbrook (bottom). The symbols denote the points (u_t, v_t) at times t=0, 6, 12, 18, 24 h where (u_0, v_0) (t=0 h) is marked by "+". The first three PC are stationary with slight wind speed variations over the day. The higher PC describe changes in the meteorological situation.*

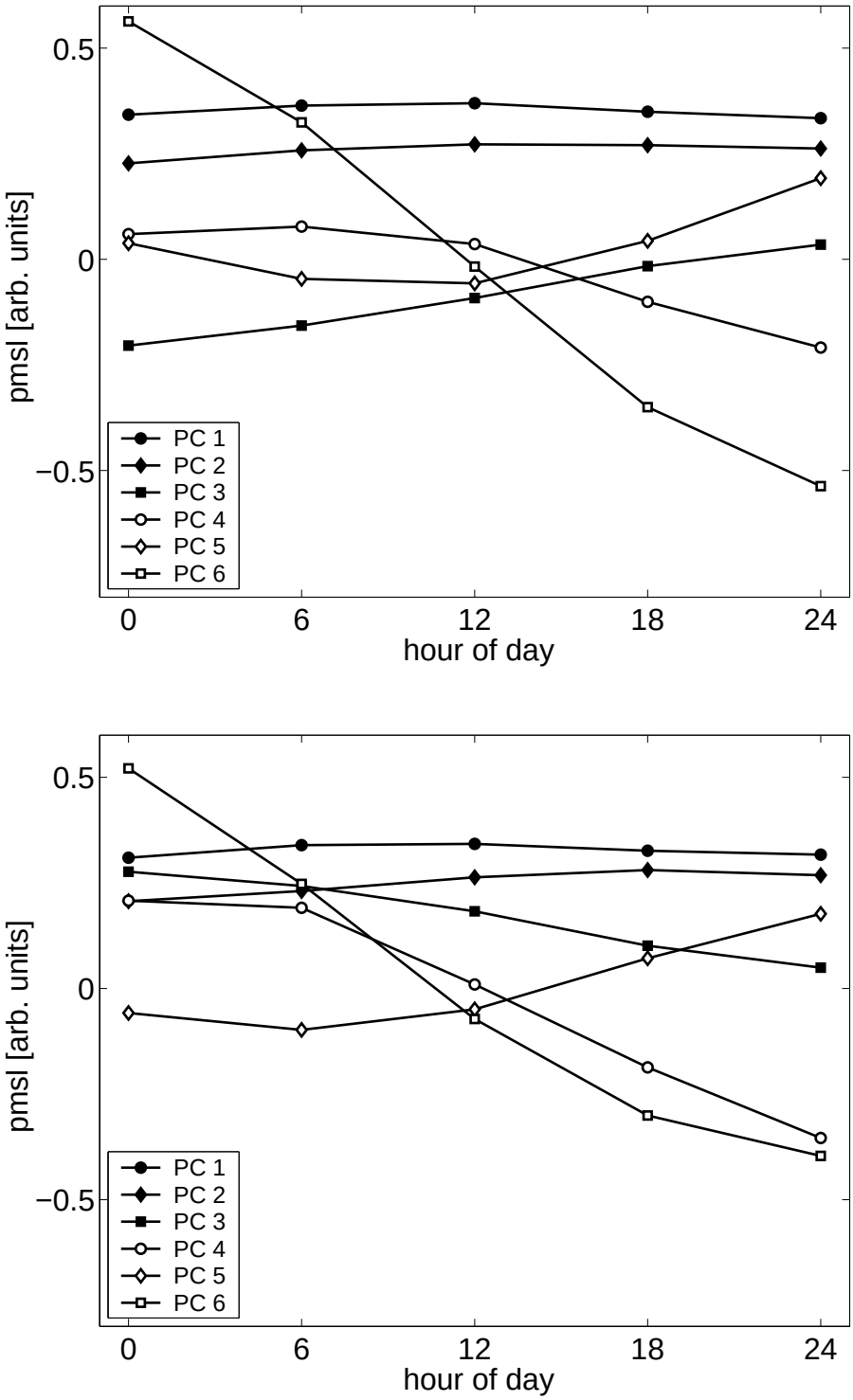

Figure 3.3: *First six PC of pmsl space for Fehmarn (top) and Hilkenbrook (bottom). The first two PC refer to rather constant high pressure while the higher PC describe changes in pmsl which can be rather dramatic, e.g. PC 6.*

ing to Equation (3.6) of the two clusters that have been formed in each step. The larger the distance the greater the difference between the clusters in terms of the chosen linkage method. Hence, the clustering process should be interrupted if the distances between the clusters become too large which is often indicated by a "jump" in the series of distances.

Figures 3.4 and 3.5 show the distances of the final 15 clustering steps using complete linkage at the two sites. In both cases there are several jumps in the distances that suggest the unification of two clusters that are quite distinct. For Fehmarn two crucial points can be identified at the transition from seven to six and from five to four clusters. At Hilkenbrook detectable steps in the cluster distances occur at the transition from eight to seven and also five to four clusters.

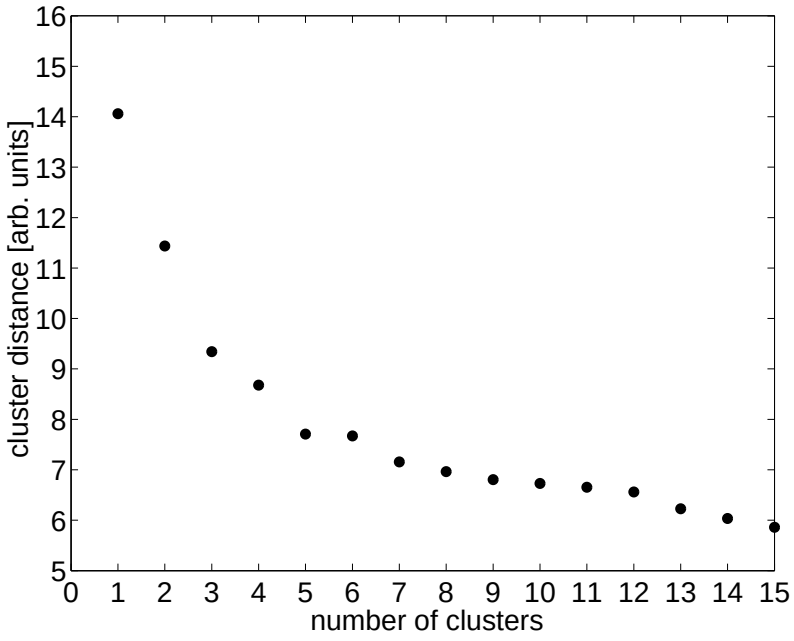

Figure 3.4: *Distance between clusters versus number of clusters using complete linkage at Fehmarn. As clusters are successively joined "jumps" in the distances start to occur at the transitions from seven to six and five to four clusters. This defines the range of cluster numbers considered for further analysis.*

In this investigation it turns out that the classification scheme is quite robust in the sense that useful classification results can be obtained for different cluster numbers. The behaviour of the cluster distances suggests a range of possible cluster numbers that should be considered for further investigations rather than providing just one optimal number. Which number of clusters is selected depends on the detailed purpose of the investigation. In order to interprete clusters in terms of weather classes and to compare clusters found at different sites it is useful to choose a number of six to eight clusters because the synoptic situations are more homogeneous. However, using more clusters means that a smaller proportion of the 365 days per year are grouped together in a common cluster. This can lead to less significant results with regard to the average forecast error. As a consequence, it seems advisable to vary the number of clusters in the range

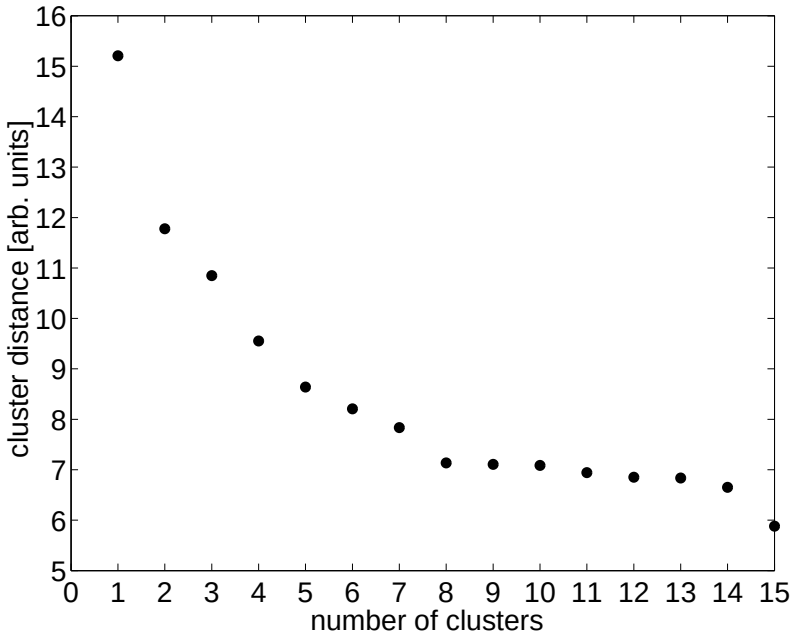

Figure 3.5: *Distance between clusters versus number of clusters using complete linkage at Hilkenbrook. Note the discontinuities at the transitions from eight to seven and five to four.*

indicated by the discontinuities in the cluster distances and to carefully consider the results in terms of the corresponding weather class on the one hand and the forecast error on the other hand. As the clustering technique has the convenient property that only two clusters are joined at each step without affecting the remaining groups it is easy two follow what actually happens if the number of clusters is varied.

Complete linkage at Fehmarn

Following this approach six clusters are used for Fehmarn to see how they can be connected to the overall weather situation. After the clusters are defined the corresponding meteorological situation is considered in terms of the mean values of "real" $\vec{u}$ and pmsl for each cluster. Figure 3.6 shows the mean values of the meteorological variables of Fehmarn's six clusters. Note that $\vec{u}$ denotes the wind vector and points in the direction in which the wind blows whereas wind directions, as usual in meteorology, refer to the direction from which the wind blows. In terms of pmsl (Figure 3.6 (bottom)) the clusters show three different regimes: low pressure (clusters 1, 2 and 5), moderately high pmsl (clusters 3 and 6) and very high pmsl (cluster 4).

The low pressure situations differ considerably with respect to wind speed and direction (Figure 3.6 (top)). Typically, cluster 1 has moderate winds from south west and slightly decreasing pressure caused by a low pressure system approaching from the west or north west together with a high south of the site. The large-scale weather map of a typical day of cluster 1 is shown in Figure 3.7 (top). Cluster 2 refers to a strong low that passes north of the site as illustrated in Figure 3.7 (bottom). The pmsl drops on average

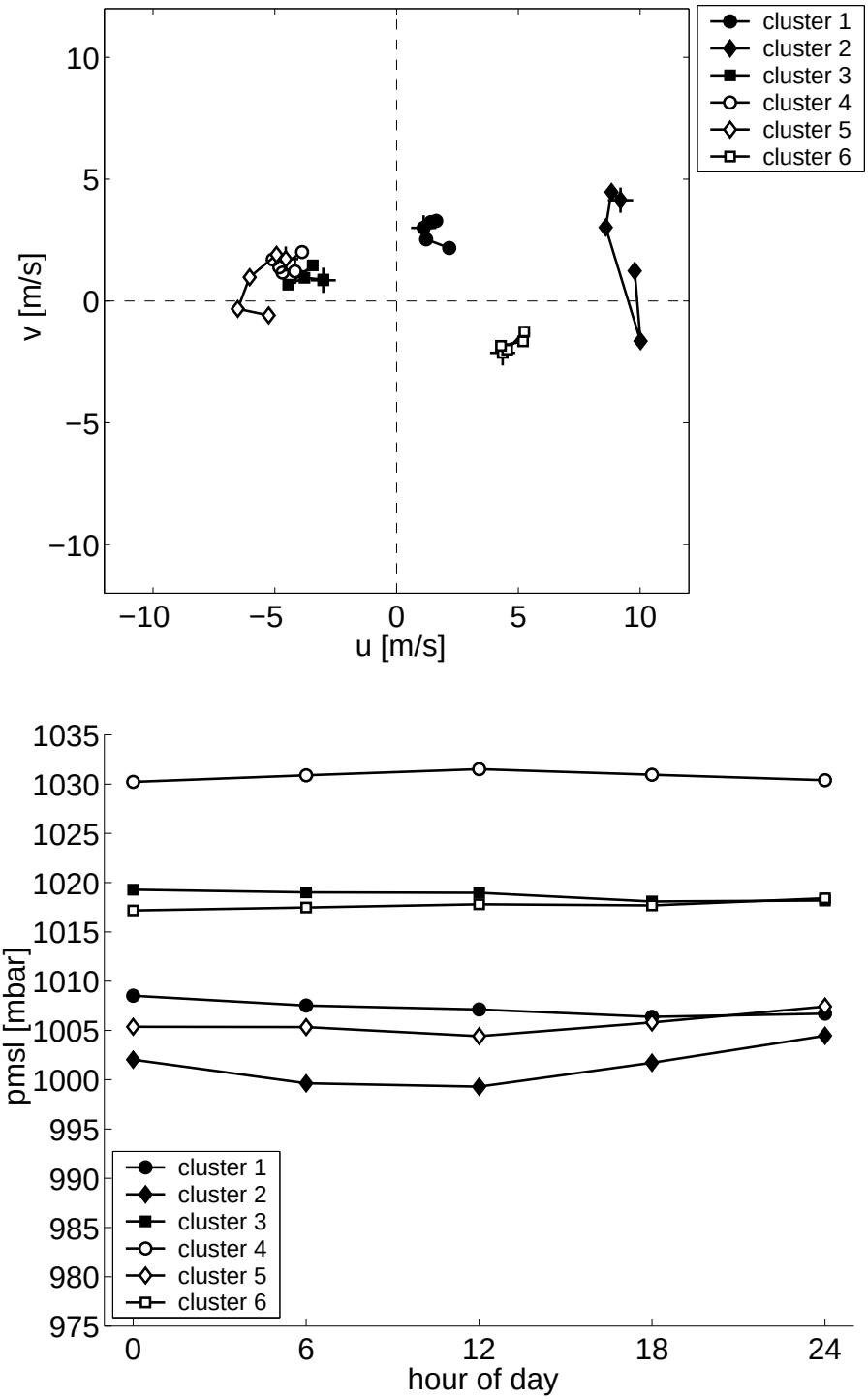

Figure 3.6: *Means of $\vec{u}$ (top) and pmsl (bottom) for different forecast horizons for six clusters at Fehmarn constructed with complete linkage. For $\vec{u}$ the symbols denote the points (u_t, v_t) at times t=0, 6, 12, 18, 24 h where (u_0, v_0) (t=0 h) is marked by "+". The corresponding weather situations are shown in Figures 3.7, 3.8 and 3.9.*

to a low level around 1000 mbar and recovers again at the end of the day. The wind speeds are very high with wind direction changing from south west to west north west and then back to south west. Cluster 5 is related to a situation where the low passes west of the site, e.g. over Britain and France, with high pressure gradients over the Baltic Sea (Figure 3.8 (top)). Wind speeds are quite considerable turning from south east to north east.

Cluster 3 is related to moderate wind speeds from easterly directions. This typically occurs if a stable high pressure area persists over western Russia and a rather strong low is located west of the site leading to considerable pressure gradients over Fehmarn (Figure 3.8 (bottom)).

Days in cluster 6 are characterised by almost the same mean pressure as in cluster 3 but with north westerly wind directions. In this case central Europe is influenced by a high or ridge located west or north west of the site (Figure 3.9 (top)).

Cluster 4 summarises mainly days from the winter period with very high pmsl around 1030 mbar and easterly winds. The corresponding weather situation is typically associated with high pressure over Scandinavia extending over Central Europe as shown in Figure 3.9 (bottom).

Comparing the large-scale weather maps of days within one cluster shows that the overall weather situation for most of the days is rather coherent. Hence, the classification scheme that is based on local data at the site can be related to the overall weather situation. This is not totally surprising as the local meteorological situation described by wind speed, direction and pressure is caused by the atmospheric circulation on a larger scale. Of course, for some days the weather situations are not very precisely described by the mean values of the cluster. If more classes are used clusters typically split into new clusters which are in itself more homogeneous than before. But with regard to the investigation in this work a small number of clusters is desired and sufficient even though the division might be too rough for other purposes. On the whole, the synoptic classification using six clusters with complete linkage for Fehmarn appears to be reasonable.

Complete linkage at Hilkenbrook

Hilkenbrook is about 250 km west of Fehmarn and, therefore, expected to be exposed to different wind conditions, in particular, less influenced by continental high pressure systems over Scandinavia and Russia but more affected by lows approaching from the Atlantic Ocean. However, there should be a number of weather situations which are comparable to Fehmarn as the size of the overall circulation patters is much larger than the distance between the sites. This is confirmed by considering the means of the meteorological variables in the seven clusters created by complete linkage for Hilkenbrook in Figure 3.10. Again the clusters are roughly ordered according to low pressure (clusters 5 and 6), moderate pmsl (clusters 3 and 4), and high pmsl (clusters 1, 2 and 7).

The two low pressure classes both refer to rather dynamic situations with pmsl varying over the day (Figure 3.10). Cluster 5 has high wind speeds from the south with pmsl dropping from about 1005 mbar to 995 mbar due to a low pressure system approaching from the north or north west. Cluster 6 is related to a strong low passing north of the site. The wind blows rather strongly and turns from south west to west and back to south

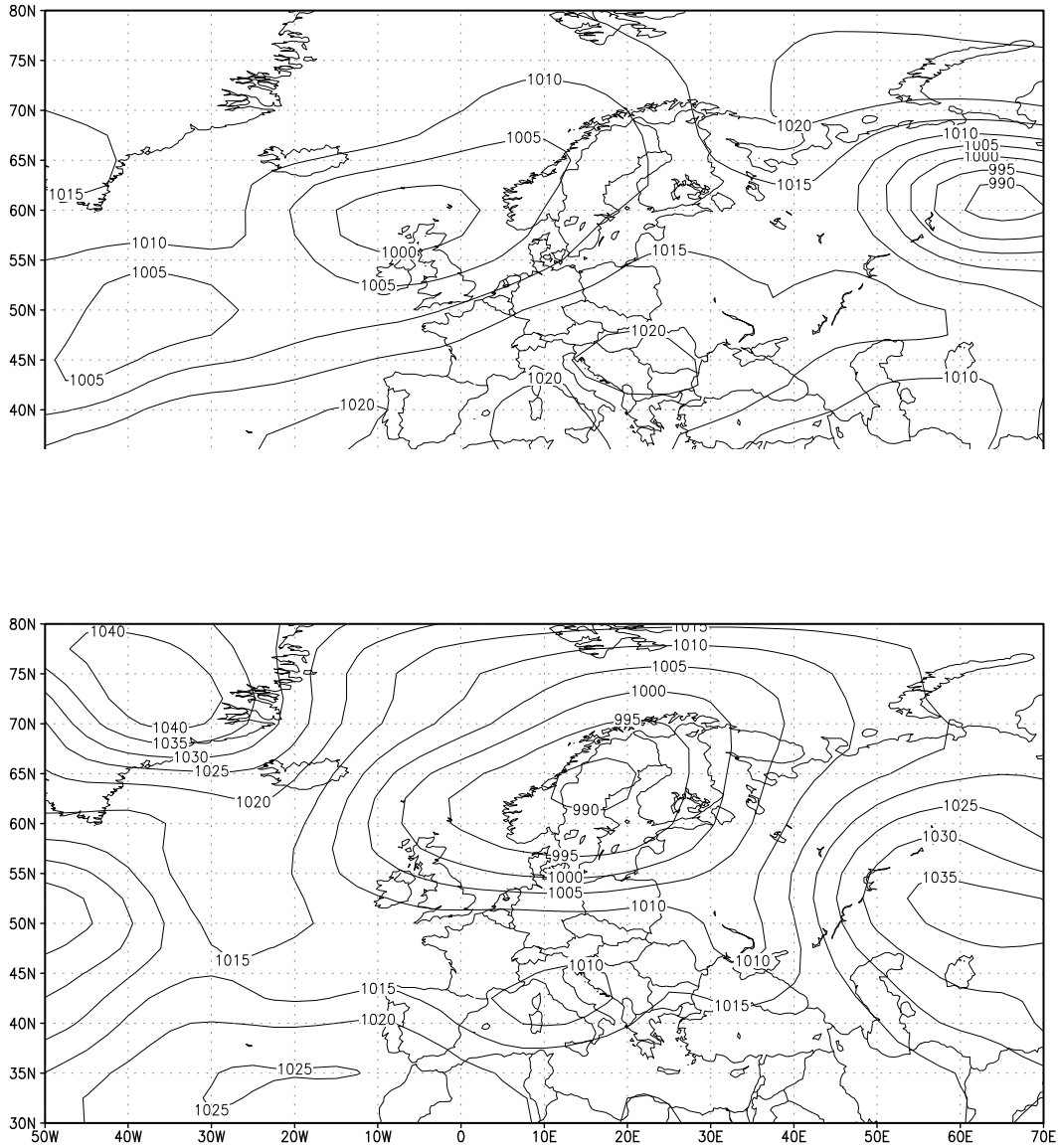

Figure 3.7: Large-scale Weather maps of days inside a specific cluster representing typical weather classes derived by complete linkage at Fehmarn located at 54 deg N, 12 deg E. Top: Low pressure approaching from North West (cluster 1). Bottom: Strong low passes north of Germany (cluster 2).

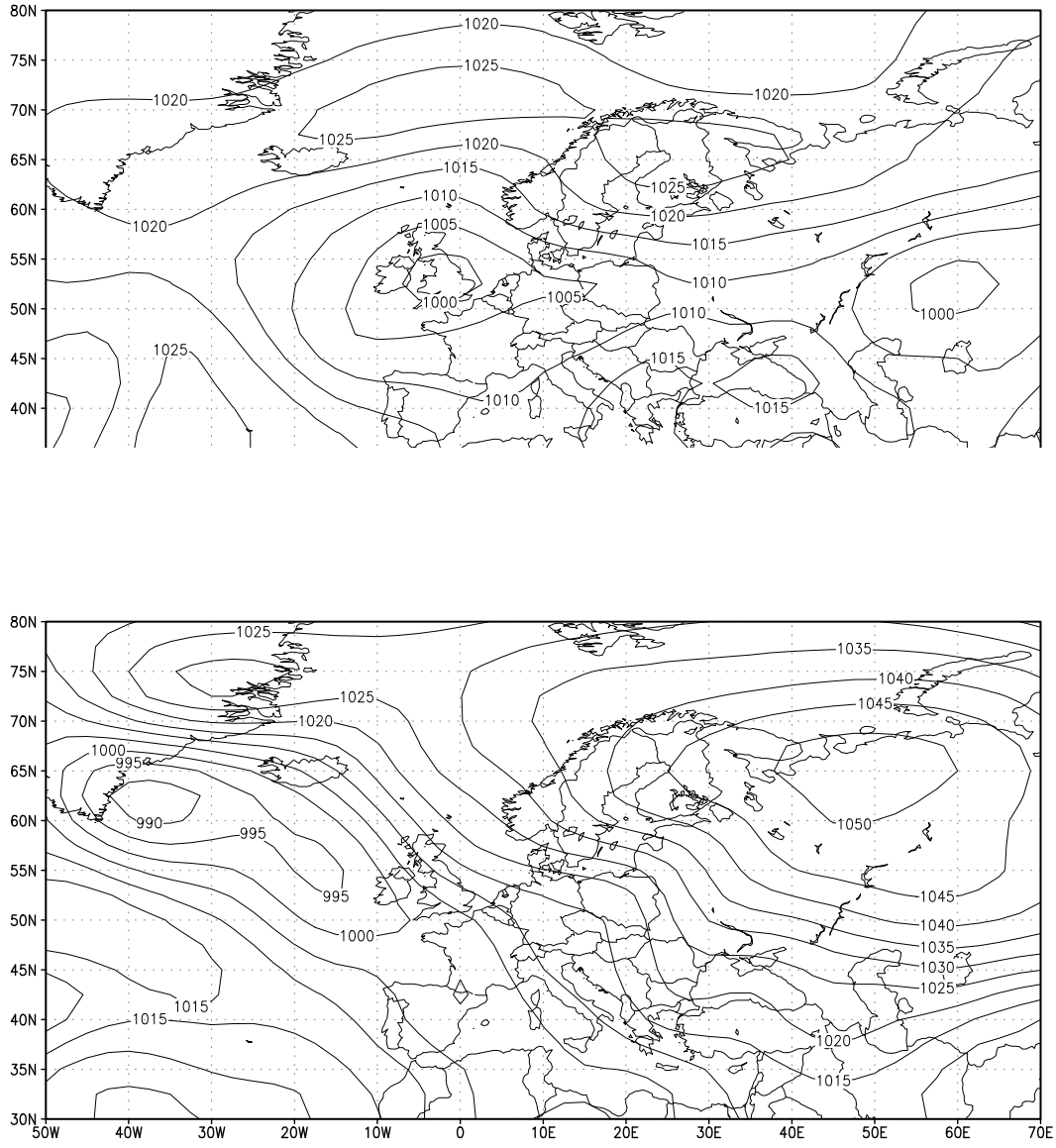

Figure 3.8: *Same as Figure 3.7 but for different weather classes. Top: Low passing west of Fehmarn over Britain and France (cluster 5). Bottom: Stationary high pressure over western Russia (cluster 3).*

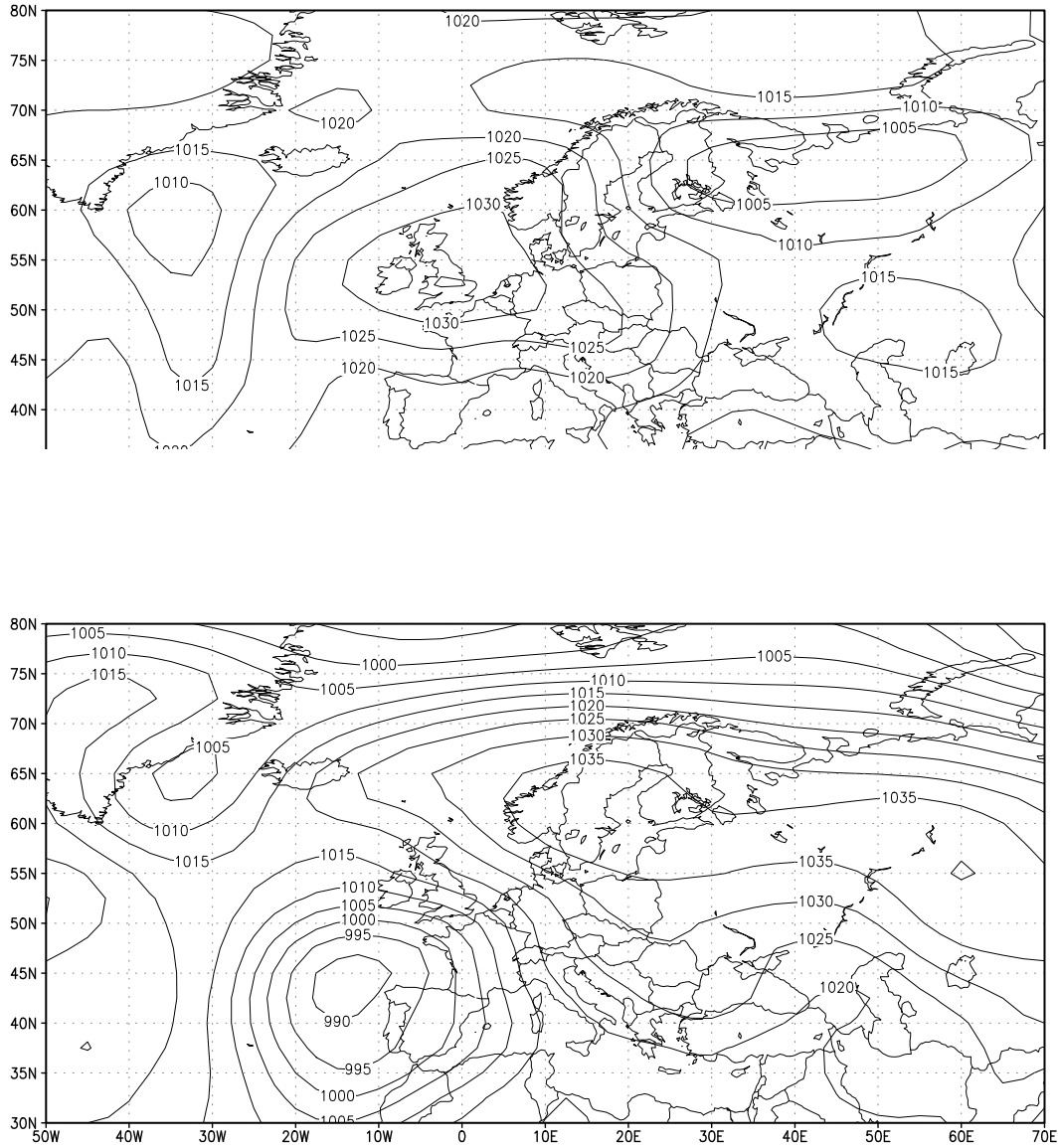

Figure 3.9: *Same as Figure 3.7 but for different weather classes. Top: High pressure area north west of Germany with low over western Russia (cluster 6). Bottom: Strong high pressure system over Scandinavia with low influencing west of Europe (cluster 4).*

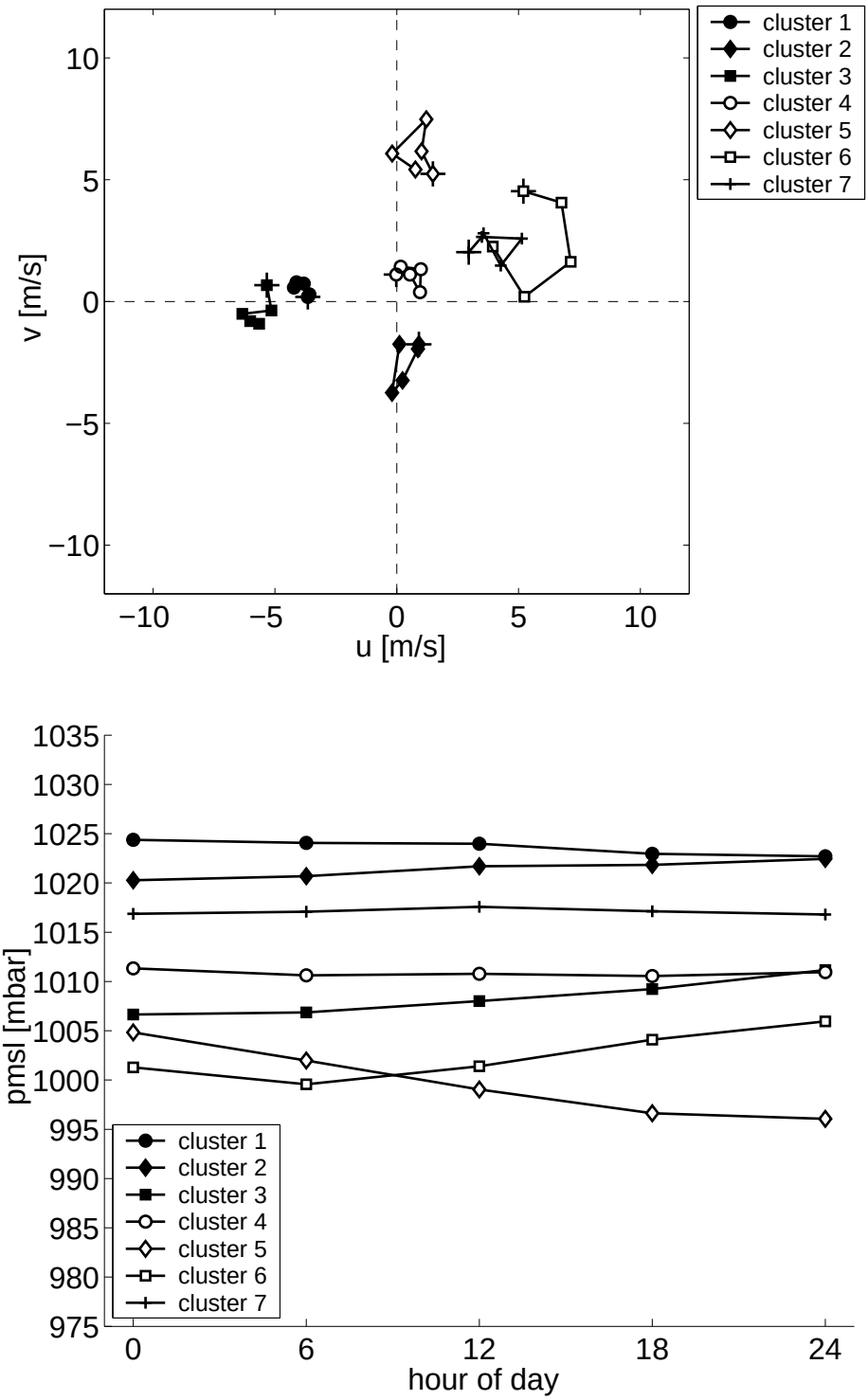

Figure 3.10: *Means of $\vec{u}$ (top) and pmsl (bottom) at different prediction times for seven clusters found for Hilkenbrook with complete linkage. For $\vec{u}$ the symbols denote the points (u_t, v_t) at times t=0, 6, 12, 18, 24 h where (u_0, v_0) (t=0 h) is marked by "+".*

west indicating the passage of a front.

Moderate pmsl around 1010 mbar occurs at Hilkenbrook mainly in two different weather regimes. Cluster 3 shows rather high wind speeds from the east and slightly increasing pressure typically related to a high pressure system east or north east of the site, e.g. over western Russia, and low pressure in the west or south west. Cluster 4 contains quite a number of days which have pmsl around 1010 mbar and very low wind speeds. The weather class is not as clear as in all the other clusters but mainly characterised by the fact that the centres of the pressure areas are far away with small pressure gradients at the location of the site.

The three high pressure clusters at Hilkenbrook show quite pronounced differences concerning their mean wind directions. Cluster 1 has on average the highest pmsl of about 1023 mbar. Note that this is more than 5 mbar lower compared to the highest pmsl at Fehmarn. The wind directions are from the east with moderate speeds. The overall weather situation is typically dominated by a stationary high pressure system over Scandinavia and, hence, north east of the site. In contrast to this, cluster 2 refers to a weather situation where the high pressure area is located in the west of the site leading to northerly wind directions. Finally, cluster 7 is related to south westerly winds caused by high pressure south west or south of the site and at the same time low pressure in the north.

Hence, also for Hilkenbrook the classification of meteorological situations based on complete linkage appears to be useful.

The typical weather classes found for the two sites seem to be rather similar in terms of the general description of the meteorological situations. In order to further test the consistency of the classification scheme the days that simultaneously appear in clusters from Fehmarn and Hilkenbrook are counted. The result is shown in table 3.1 where the rows refer to Fehmarn's clusters and the columns to those from Hilkenbrook.

Table 3.1: *Comparison of equal days in clusters for Fehmarn and Hilkenbrook.*

	Hilkenbrook							
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	$\sum$
Fehmarn								
cluster 1	0	0	0	29	12	6	3	50
cluster 2	0	0	0	0	0	15	2	17
cluster 3	30	16	15	51	2	0	7	121
cluster 4	35	0	0	0	0	0	0	35
cluster 5	0	0	3	16	1	1	0	21
cluster 6	4	43	0	38	0	0	36	121
$\sum$	69	59	18	134	15	22	48	365

Some clusters share a rather profound number of days. For example, cluster 2 of Fehmarn is, except of two days, completely contained in Hilkenbrook's cluster 6. Hence, on these 15 days the classification at both sites records a weather situation dominated by a strong low pressure system that is centred north of Germany and moves quickly to the north east. However, the exact path and the advection time of the low decides

whether both sites are affected simultaneously or only one of them. If the low takes a more southern route only Hilkenbrook classifies this situation as cluster 6 while Fehmarn has a smaller pressure drop and lower wind speeds as it is located further in the east and, hence, might not to affected by the fronts. Fehmarn labels this situation as cluster 1.

Another example is cluster 4 of Fehmarn which is totally absorbed by cluster 1 of Hilkenbrook. These 35 days are typically related to a high pressure system over Scandinavia or Western Russia with extremely high pmsl over north eastern Germany. However, Hilkenbrook's cluster 1 contains more days than that which overlap with cluster 3 of Fehmarn. At these days the high pressure system is located further in the east leading to similar wind directions from the east and pmsl around 1020 mbar at both sites.

On the other hand, certain weather situations lead to different classification results at both stations, e.g. cluster 3 of Fehmarn is distributed over six different clusters of Hilkenbrook. This occurs if the two sites are influenced by distinct weather regimes. In this case Fehmarn is dominated by rather high pressure located over western Russia while at the same time Hilkenbrook can experience a variety of high or lower pressure situations.

The result of this comparison is that the consistency between the classifications at the two sites seems to be rather high. For the majority of cases common days in clusters can be plausibly explained by the overall weather situation that either shows that the two sites are affected in the same way or why they simultaneously record different local conditions.

Relation to forecast error using complete linkage

For each of the clusters the mean forecast error is determined by averaging the rmse values calculated for each day with Equation (3.9). The wind speed prediction at 10 m is provided by the "Deutschlandmodell" of the German Weather Service while the corresponding wind speed measurements are from the WMEP measurement programme.

The means of the daily rmse values per cluster normalised to the annual average of the wind speed are shown in Figure 3.11 for Fehmarn and Figure 3.12 for Hilkenbrook found with complete linkage. In both cases there are considerable differences between the forecast errors of the different clusters. The error bars illustrate the 95% confidence intervals of the mean values suggesting that the clusters with minimum and maximum rmse are indeed significantly separated.

In particular, Fehmarn's clusters 1, 2 and 5 are related to large relative rmse values around 0.38 corresponding to about 2.1 m/s absolute rmse. As described before (Figure 3.6) these three clusters are related to different low pressure situations. Cluster 2 corresponds to situations where a strong low passes north of the site and has on average the highest forecast error. In contrast to this cluster 4 representing the weather type with the largest pressure has the smallest rmse of about 0.23 relative and 1.3 m/s absolute. A statistical F-test (with confidence level 0.05) confirms that this is significantly lower compared to the above mentioned classes 1, 2 and 5. The ratio between the largest and the smallest rmse is 1.7 which is very profound.

At Hilkenbrook cluster 6 has the highest rmse with 1.8 m/s absolute and 0.51 relative. The corresponding weather situation is related to a low pressure system passing north of

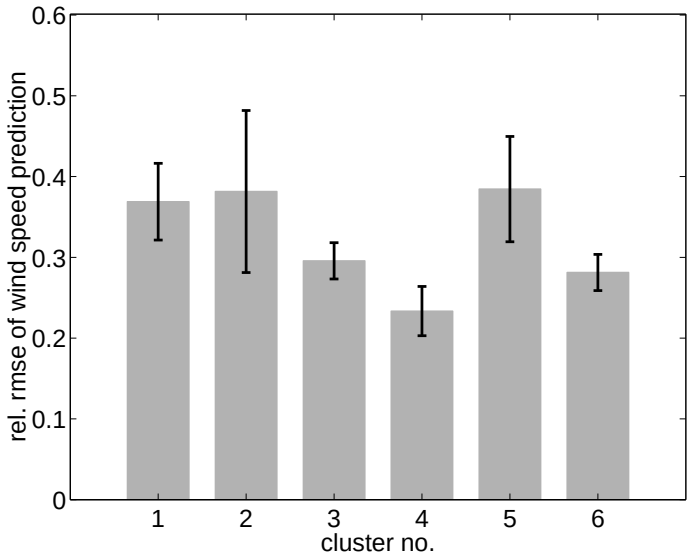

Figure 3.11: *Means of daily rmse of wind speed prediction normalised with annual mean of wind speed for the six clusters found with complete linkage at the site Fehmarn. The error bars illustrate the 95% confidence intervals of the mean. Clusters 1, 2 and 5 show the largest error. In both cases the corresponding weather situation is dominated by a low pressure system. All three clusters have significantly larger forecast errors than cluster 4 which typically refers to a stationary high pressure situation and has a significantly lower rmse compared to clusters 1, 2 and 5.*

the site. The cluster with the smallest average forecast error is cluster 1 related to a rather stationary high pressure situation where the rmse is 1.0 m/s absolute and 0.29 relative to the mean wind speed. The ratio between the maximum and minimum rmse is 1.8. It is also interesting to note that the forecast error in cluster 4 is on a medium level of 0.39 relative rmse (1.4 m/s) which is roughly half way in between the smallest and largest error values and significantly different from both of them. The days in this cluster are related to medium pmsl with small pressure gradients and, hence, low wind speeds.

Again the average forecast errors found for the typical weather situations at the two sites show a good consistency. For Fehmarn as well as for Hilkenbrook large rmse values occur for the clusters that are related to a rather fast moving low pressure system that passes north of the site (cluster 2 for Fehmarn, 6 for Hilkenbrook). The local meteorological conditions are characterised by pmsl on a relatively low level further decreasing around midday but recovering at the end of the day. Wind speeds are fairly high with wind directions turning from south west to west and back to south west indicating the passage of a front. Depending on the path of the low this situation simultaneously occurs at both sites or only at Hilkenbrook (see table 3.1).

In addition, for Fehmarn the passage of low pressure areas in the west or south west of the site with strong easterly winds (cluster 5) is also related to large forecast errors. At Hilkenbrook these days are recorded in cluster 4 which additionally contains more situations with on average smaller wind speeds and slightly higher pmsl compared to Fehmarn. The corresponding forecast error is the second largest at Hilkenbrook but sig-

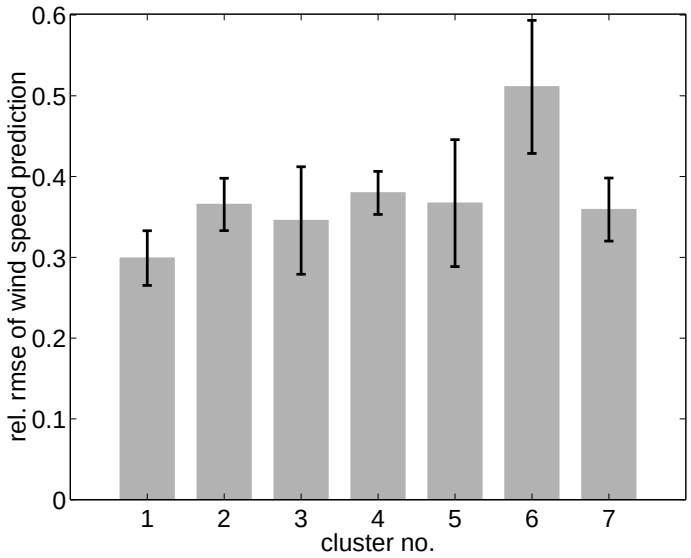

Figure 3.12: *Same as Figure 3.11 but for the seven clusters found with complete linkage at Hilkenbrook. Cluster 6 has the maximum forecast error which is significantly different from clusters 1, 2, 4 and 7. Cluster 6 is related to a low pressure system passing north of the site. Cluster 1 with the smallest rmse corresponds to a situation where a stationary high pressure dominates the local weather conditions. Its error is significantly lower compared to clusters 4 and 6.*

nificantly lower than the largest one. Thus, at Fehmarn there are higher wind speeds caused by larger pressure gradients while at the same time at Hilkenbrook more moderate conditions prevail.

In contrast to this, both sites have the smallest forecast error in situations where a high pressure area lies rather stationary over Scandinavia or the Baltic Sea with relatively high wind speeds from the east. For Fehmarn the mean rmse in this case is 17% to 39% smaller than those of the other weather classes. For Hilkenbrook the minimum rmse is 13% to 42% smaller than in the other clusters.

Hence, using complete linkage leads to a reasonable synoptic classification of the meteorological data (30 m wind data and pmsl of nearby synop station) at these two sites in the sense that the clusters can be associated with typical large-scale weather classes. The number of classes found by considering the interval of clustering steps given by "jumps" in the distances between the clusters as criterion to stop the clustering process already provides a good choice. Moreover, the mean forecast errors related to the clusters show significant differences where the maximum and minimum rmse are as expected related to dynamic low pressure and stationary high pressure, respectively.

Ward's linkage

For most of the investigated sites Ward's linkage produced clusters which are in principle comparable to those found by complete linkage. But the alterations between the groups of days generated by the two methods can make important differences in terms

of the grouping of days and, thus, the related forecast errors. This is now illustrated for Fehmarn where typical differences between the methods occur.

As Figure 3.13 shows the distances between the clusters at the final 15 steps of the clustering procedure behave rather similar to complete linkage. There are discontinuities at the transitions from eight to seven and five to four clusters. Again choosing a number of clusters from this interval leads to a classification that allows to relate the clusters to large-scale weather situations and to find significant differences between the average forecast errors per cluster.

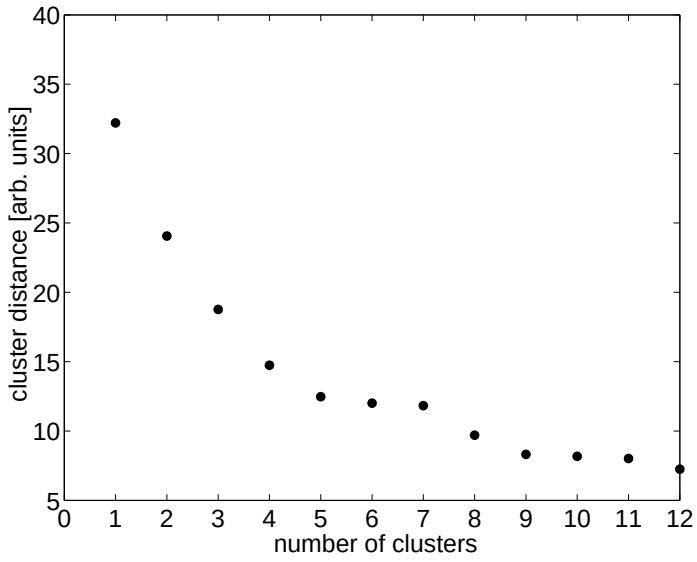

Figure 3.13: *Distance between clusters versus number of clusters using Ward's linkage at Fehmarn. "Jumps" in the distances occur at the transitions from eight to seven and five to four clusters. This is similar to complete linkage (Figure3.4).*

The results of the classification based on six clusters with Ward's linkage for Fehmarn are shown in Figures 3.14. With regard to these mean values of $\vec{u}$ and pmsl the classes are generally comparable to those found by complete linkage in Figures 3.6. But it is obvious by comparing the means of the pressure that there are differences in particular for the low to medium range with pmsl between 1005 mbar and 1015 mbar. Moreover, in terms of the mean wind vectors Ward's cluster 1 appears to constitute a new class which does not appear under complete linkage.

A more detailed comparison of the overlaps between the clusters reveals that some days are grouped rather differently by the two methods. Table 3.2 shows that the only two clusters that are nearly identical are number 2 of complete and 5 of Ward's which share 16 common days related to the well-known weather situation of the passing low in the north. Hence, these seldom but rather extreme days are consistently classified by both methods.

Another interesting weather class is the high over Scandinavia or the Baltic Sea detected by complete linkage as cluster 4. Under Ward's linkage these 33 days with very high pressure are almost totally contained in cluster 2. But this cluster contains addi-

tional days from situations with slightly lower pmsl and lower wind speeds. Hence, Ward's cluster 2 joins days into one class that are separately classified under complete linkage, i.e. clusters 3 (pmsl around 1020 mbar) and 4 (very high pmsl $\approx$ 1030 mbar). This tendency of Ward's linkage to group extreme together with less extreme situations though they should remain distinct has already been described in earlier investigations, e.g. by Kalkstein et al. [46]. Another typical feature of Ward's method is to create rather equally sized groups which can also be observed in table 3.2. The distribution of days on the clusters is more balanced compared to complete linkage.

Table 3.2: *Comparison of equal days in clusters complete and Ward's linkage for Fehmarn. Though the means of the clusters are rather similar to complete linkage there are differences in the actual grouping.*

	Ward						
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	$\sum$
Complete							
cluster 1	0	0	30	0	4	16	50
cluster 2	0	0	1	0	16	0	17
cluster 3	10	24	32	0	0	55	121
cluster 4	0	33	0	2	0	0	35
cluster 5	0	0	0	0	1	20	21
cluster 6	22	0	41	55	3	0	121
$\sum$	32	57	104	57	24	91	365

Relation to forecast error using Ward's linkage

The mean daily forecast error per cluster for Ward's linkage is shown in Figure 3.15. Compared to the rmse of the complete linkage clusters in Figure 3.11 the differences between the classes are less pronounced. The statistical F-test reveals that cluster 2 with the lowest rmse is significantly different from clusters 3 and 6 with the highest rmse. But in contrast to this, complete linkage results in a higher number of significantly different clusters, in particular the cluster with the smallest prediction error has a significantly lower rmse than three of the other clusters. Thus, at this site Ward's clusters have rather equalised error levels with smaller differences between the clusters. This is considered as a disadvantage compared to complete linkage as the classification scheme that provides a better distinction between the forecast errors is more useful. Of course, under the condition that it produces rather homogeneous clusters which can be associated with certain weather types.

For the sites in this investigation the clusters constructed by Ward's linkage provided a classification scheme that appears to correspond to typical weather situations. But compared to complete linkage the grouping of days can be different with a slight trend to join distinct meteorological situations into common classes and to form equally sized groups. However, in contrast to the findings by Kalkstein et al. [46] Ward's linkage does not appear "to blur distinctions between types". In terms of the related prediction error complete linkage performs better than Ward's linkage at Fehmarn and Hilkenbrook as it

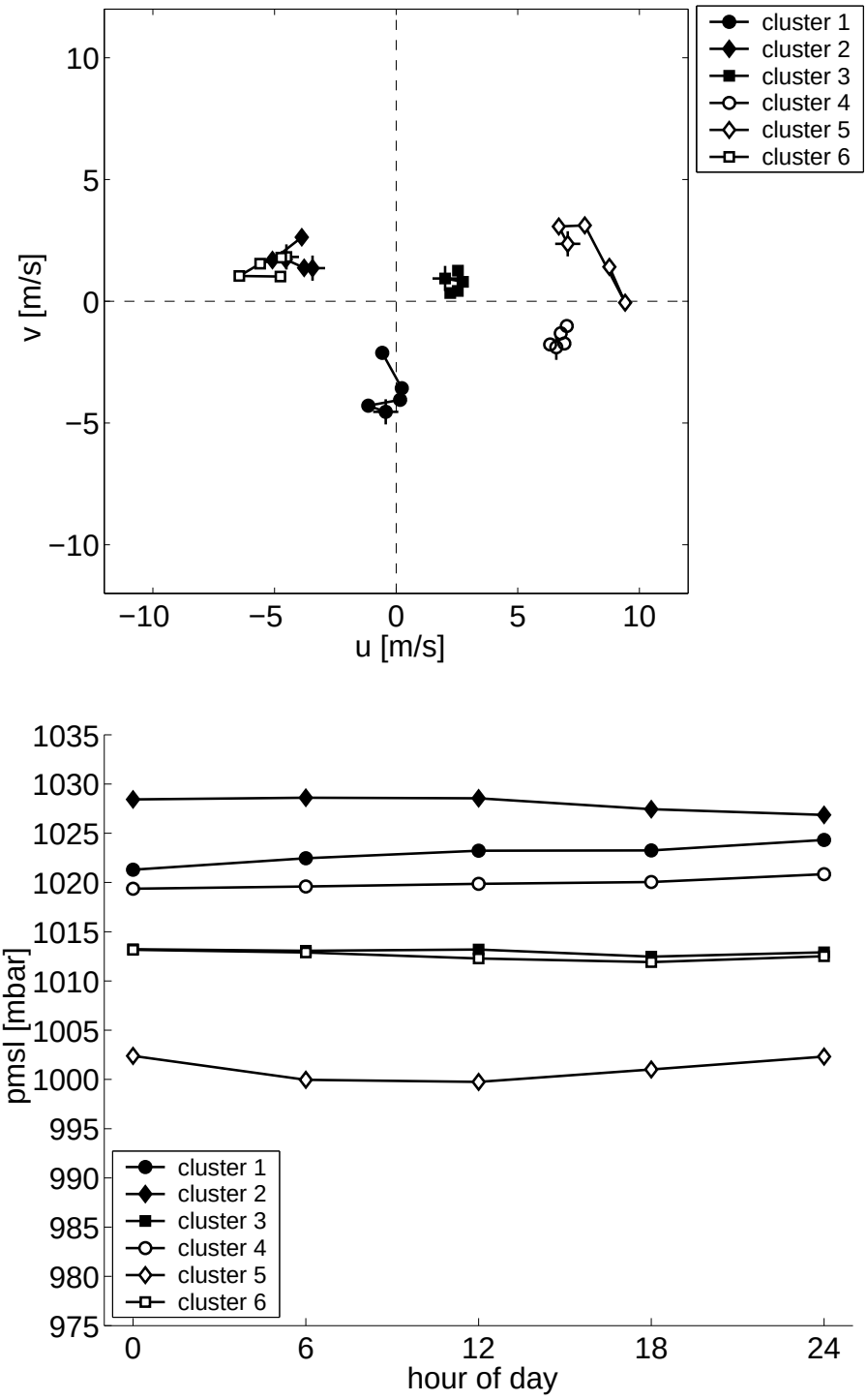

Figure 3.14: *Means of $\vec{u}$ (top) and pmsl (bottom) at different forecast horizons for six clusters found by Ward's linkage at Fehmarn. For $\vec{u}$ the symbols denote the points (u_t, v_t) at times t=0, 6, 12, 18, 24 h where (u_0, v_0) (t=0 h) is marked by "+". Some of the clusters are comparable with those provided by complete linkage (Figure 3.6), in particular cluster 5 with cluster 2 by complete linkage.*

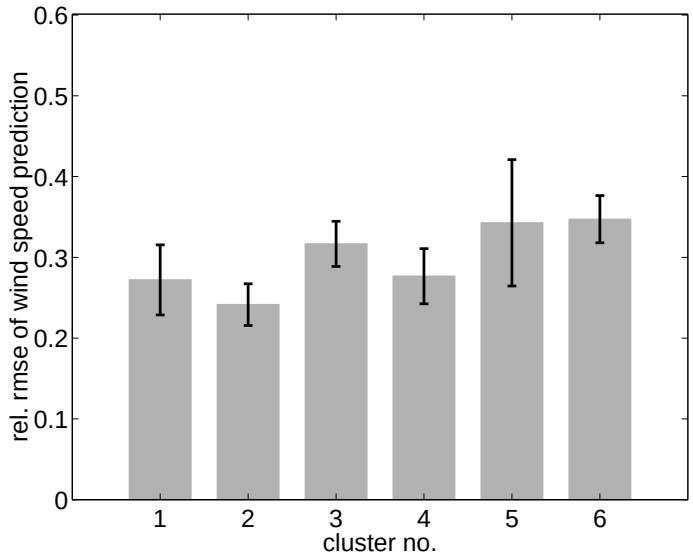

Figure 3.15: *Means of daily rmse for the six clusters found with Ward's linkage at Fehmarn. The differences between the forecast errors are not as pronounced as for complete linkage (Figure 3.11).*

provides sharper distinctions between the clusters.

Results for other sites

The site Syke is about 50 km west of Hilkenbrook still in the north western part of Germany. The results based on complete linkage using of measurements of wind data at 10 m height instead of 30 m as before are rather similar to Hilkenbrook. Here the passage of the low centred in the north is also related to the maximum forecast error. The minimum rmse also occurs for a high pressure situation but in contrast to the two sites above the high is west of the site. The Scandinavian high that leads to the minimum forecast errors for the sites above still has a relatively small rmse. At Syke the performance of Ward's technique is comparable to the complete method. The behaviour of the forecast error for the different weather classes, particularly, the systematically larger rmse for dynamic low pressure situations is also confirmed under Ward's linkage.

At Rapshagen complete linkage (also 10 m wind data) provides clusters whose interpretation in terms of large-scale weather classes is not as clear as for the other sites. At this site Ward's linkage seems to be superior and leads to a classification that is easier to relate to weather patterns. The forecast errors per cluster again show that the well-known low pressure passage has a rather high rmse but certain high pressure conditions also have. However, this site should be treated with care as local effects seem to strongly influence the flow conditions locally though the orography is only slightly complex. In terms of the relation between meteorological conditions and the forecast error this leads to an overlap between the overall predictability of the weather situation by the NWP and local effects on the wind field due to the specific on-site conditions.

Regarding local effects, e.g. due to orography causing speed up effects over hills and

channeling effects in valleys, it has to be points out that the classification method works with on-site information and is, therefore, unable to tell local flow distortion from global circulation. Hence, local effects are implicitly taken into account and there might be an overlap between the two effects. Among the investigated sites only one site (Rapshagen) obviously shows this phenomenom and makes it difficult to draw definite conclusions while the other sites do not show strong signs of local effects in the course of the analysis.

As a summary the results for all investigated sites are given in table 3.3 where the relative rmse for the three weather classes "low passes in the north", "Scandinavian high" and "high centred in north west" having the most distinct differences in terms of the forecast error are shown. The rather dynamic passage of the low leads to large prediction errors at all sites with the relative rmse ranging from 0.31 to 0.51. Compared to this the two high pressure situations are related to significantly smaller forecast errors. The high over Scandinavia causes an average rmse from 0.23 to 0.29 with the exception of the unusual error value of 0.43 at Rapshagen presumably related to orographic effects. A high pressure area centred north west or west of the sites has also rather small relative rmse between 0.22 and 0.36. To express the difference between the weather classes in terms of the forecast error the ratio between the maximum rmse of the low pressure situation and minimum rmse of the high pressure is used. It ranges from 1.5 to 1.7 where again Rapshagen is not considered. Hence, the differences are rather profound and the weather type appears to be an important criterion to distinguish different error regimes.

Table 3.3: *Overview of the relation between the average forecast error and the meteorological situation for the investigated sites using complete linkage for Fehmarn, Hilkenbrook and Syke and Ward's linkage for Rapshagen. The number of clusters is determined by inspection of cluster distances. The relative rmse of those weather classes are shown that typically have high ("low passes in the north") or low ("Scandinavian high", "high pressure north west") forecast errors. At Rapshagen the flow situation corresponding to high pressure with easterly winds seems to be influenced by local effects and produces a very unusual large forecast error marked by $*$.*

	number of clusters	relative rmse "Low passes " north of site"	relative rmse "Scandinavian high"	relative rmse "High centred in north west"	rmse ratio max./min.
Fehmarn	6	0.39	0.23	0.29	1.7
Hilkenbrook	7	0.51	0.29	0.36	1.8
Rapshagen	7	0.31	0.43*	0.22	2.0*
Syke	7	0.36	0.29	0.24	1.5

3.4 Conclusion

In this chapter the quantitative relation between the actual weather situation and the error of the corresponding wind speed prediction has been investigated with methods from synoptic climatology using local measurements of meteorological variables. The main result is that in this framework significant differences in the forecast error for distinct weather situations can be observed. In particular, the expectation can be confirmed

that dynamic low pressure situations with fronts are related to considerably larger prediction errors than weather types that are mainly influenced by rather stationary high pressure systems. The ratio between the maximum and the minimum error in terms of the average rmse for these situations is between 1.5 and 1.7 which is quite profound.

A classification scheme based on principal component analysis and cluster analysis is successfully applied to automatically divide meteorological situations into different classes. Although the classification procedure uses local information of wind speed, wind direction and atmospheric pressure the derived classes can be associated with the overall weather situation in most cases by comparing typical days from the clusters with large-scale weather maps. It is important to include information about changing weather conditions by using several measurements of the variables a day. The classification results for the different sites that have been investigated are very consistent.

However, care has to be taken if local effects at the site have a strong influence on the flow, e.g. due to orography causing speed up effects over hills and channeling effects in valleys. As the classification method works with on-site information it is unable to tell local flow distortion from global circulation. Hence, local effects are implicitly taken into account and the results may reflect an overlap between the two effects. In this investigation the classification procedure seems to be robust enough to deal with the degree of local inhomogeneity that occurs for sites in northern Germany including an island site. To exclude local effects it is advisable to prefer wind data measured at 30 m height or higher to 10 m data. For further use of this type of classification scheme it is necessary to systematically evaluate the performance for sites in more complex terrain.

Moreover, it is important to note that this investigation only confirms that there is a relation between the prevailing weather situation and the forecast error of the wind speed for historical data of one year. So far, only the 00 UTC prediction run with the lead times 6, 12, 18 and 24 h has been used to assess the daily forecast error because of the limited availability of high quality data in all variables at all times. Hence, it is desirable for future investigations to shed some light on the behaviour of the prediction horizons beyond 24 hours and to confirm the results found here with data from longer periods of time.

With regard to practical applications the advantage of this method is clearly that it works on a rather small set of standard meteorological variables such that on-line measurements can be obtained quite easily and cost effectively. This is, together with the fact that the classification scheme is automatic rather than manual, a major pre-requisite for a possible operational use of the classification scheme.

So far measurements of the meteorological variables have been used to determine the weather class. In order to exploit these findings for an estimation of the uncertainty of a wind power prediction basically two steps are necessary. First of all the predictability of the weather classes themselves has to be evaluated. For prediction purposes it is required to determine the weather type in advance in terms of the predicted wind speed, wind direction and atmospheric pressure. As the forecast quality of pressure and wind direction (e.g. shown by Mönnich [67]) is considerably better than that of wind speed the prospects are quite good to accurately predict the meteorological class. This step is nontrivial because the uncertainty of the wind speed prediction which should be pro-

vided dependent on the weather class is also involved in predicting this weather class. Hence, it must be shown that there is still some advantage in doing so. In a second step the uncertainty of the wind speed forecast has to be transferred to the power forecast. The simplest approach in this direction is to consider the error propagation as applied in section 2.5 where it was shown that the power uncertainty can rather well be estimated by the product of wind speed uncertainty and the derivative of the power curve. The innovative step in terms of the results of this chapter would then be to replace the constant wind speed uncertainties by weather type dependent ones.

Chapter 4

Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts

Abstract

For operational planning it is important to provide information about the situation dependent uncertainty of a wind power forecast. Factors which influence the uncertainty of a wind power forecast include the predictability of the actual meteorological situation, the level of the predicted wind speed (due to the non-linearity of the power curve), and the forecast horizon. With respect to the predictability of the actual meteorological situation we consider a number of explanatory variables, some inspired by the literature. The chapter contains an overview of related work within the field. We consider an existing wind power forecasting system (Zephyr/WPPT) and show how analysis of the forecast error can be used to build a model of the quantiles of the forecast error. Only explanatory variables, or indexes, which are predictable are considered, whereby the model obtained can be used for providing situation dependent information regarding the uncertainty. Finally, the chapter contains directions enabling the reader to replicate the methods and thereby extend other forecast systems with situation-dependent information on uncertainty.

4.1 Introduction

In recent years a growing interest in information about the uncertainty of wind power forecasts in different weather situations has emerged. Based on wind speed measurements and standard meteorological forecasts [9] estimates the power curve of a small wind farm and then models the relation between the actual and forecasted wind speed both with respect to the mean and the covariance. Considering the same wind farm [10] use local linear quantile regression to obtain a probabilistic model based on meteorological forecasts and on observations of power. [80, 79] use consecutive forecasts and based on these a quantity called "Meteo-Risk Index" is defined. This quantity measures the agreement between the consecutive forecasts and is used to predict the uncertainty of the wind power forecast. [59] identify relations between typical weather situations and the magnitude of the forecast error. In a research project carried out together with Eltra (the

TSO in western Denmark) [71] developed a stochastic model of the forecast errors when using WPPT, Version 2 [76]. The model describes the variance and the correlation, within and between the daily forecasts.

Over the last decade much effort has been spent on developing wind power forecasting systems supplying a point forecast of the wind power production of a farm or a region. For this reason it is desirable to be able to extend existing point forecast systems to probabilistic forecast systems. In this chapter we consider forecast errors from an existing system (WPPT, Version 4, [75, 72]) and use linear quantile regression [49] together with spline bases [21] in order to obtain a model for the 25% and the 75% quantiles of the forecast errors. The methods are easily applicable and can be applied to any forecasting system using the free software called "R" (homepage http://www.r-project.org) with the add-on package "quantreg", which can be downloaded from the same homepage. The meteorological forecasts used by WPPT in the particular setup considered is the wind speed and direction 10 meter above ground level (10m a.g.l.) from DMI-HIRLAM [84] and we consider a few other forecasted variables from DMI-HIRLAM as candidates for the quantile model. Furthermore, for each of these variables, we consider risk-indices inspired by the Meteo-Risk Index mentioned above. It was decided to focus on the horizons relevant for reporting to NordPool (http://www.nordpool.com). Considering timing and calculation times we have therefore focused on the 06Z DMI-HIRLAM forecast for horizons 18–42 hours. However, in order to be able to calculate the risk indices for each of the meteorological variables we consider only 18–36 hours, see the beginning of Section 4.4.2.

The outline of the chapter is as follows. The data, including training and test periods, is described in Section 4.2. The methods used in the chapter, i.e. quantile regression and parametric additive models, are briefly described in Section 4.3. Also, in Section 4.3, it is described why we have chosen to use additive models. Sections 4.4 and 4.5 describe the model building process and the evaluation on test data, respectively. Finally, in Section 4.6 we conclude on the chapter. In Section A it is outlined how the models can be fitted, visualized, and how forecasts can be produced using "R".

4.2 Data

The data used in this study consists of:

- 15 min. power averages from the Tunø Knob offshore wind farm consisting of 10 Vestas V39 turbines (500kW nominal). Location: 55° 58′ 08″ N, 10° 21′ 10″ E.
- Forecasts of the wind power production of the farm based on WPPT, Version 4. Time step 15 min., with a maximum horizon of 48 hours.
- Meteorological forecasts of air density, friction velocity, 10m wind speed and direction from DMI-HIRLAM [84]. Time step 60 min., with a maximum horizon of 48 hours.
- The period considered is January 1 – October 31, 2003. Data until June 1 is used for developing the quantile models. Data is available back to July 1, 1999. Data from

before January 1, 2003 is only used to indicate the overall spread of historic power productions.

Interpolation is used to obtain meteorological forecasts for all time points at which power forecasts and observations are available. It is noted that on Sept. 2, 2003 a model change was introduced into DMI-HIRLAM which is expected to have large influence on the forecasted 10m wind. Over time, WPPT will adapt to this change and will use the meteorological forecasts in an optimal way within the framework of the system. However, the distributional properties of the error may change permanently.

Figure 4.1 shows the observed power plotted against the forecasted power for the training and test data and for the data split at Sept. 2, 2003. The plots of the training and test data differ qualitatively. Especially, the saturation at high levels of the production occurs much more frequently in the training data than in the test data. The other plots on the Figure indicate that this could be related to the change in DMI-HIRLAM on Sept. 2, 2003. For this reason the evaluation on the test data will be performed on the total test set and on the test set split on the date just mentioned.

4.3 Methods

4.3.1 Quantile regression

Considering a random variable Y, the median is the most well known quantile and is characterized as the value $Q(1/2)$ for which the probability of obtaining values of Y above or below $Q(1/2)$ both equals $1/2$. Generally, $Q(\tau)$ is defined as the value for which the probability of obtaining values of Y below $Q(\tau)$ is τ. In quantile regression [49, 50] $Q(\tau)$, $0 < \tau < 1$, is expressed as a linear combination of some known regressors and unknown coefficients, exactly as the mean is modelled in (multiple) linear regression. Thus, the τ-quantile is modelled as

$$Q(\tau) = \beta_0(\tau) + \beta_1(\tau)x_1 + \ldots + \beta_p(\tau)x_p \,, \tag{4.1}$$

where $x_.$ are the p known regressors also called explanatory variables and $\beta_.(\tau)$ are unknown coefficients, depending on τ, to be determined from observations $(y_i, x_{i,1}, \ldots, x_{i,p})$; $i = 1, \ldots, N$.

Given the check function

$$\rho_\tau(e) = \begin{cases} \tau e & , \quad e \geq 0 \\ (\tau - 1)e & , \quad e < 0 \end{cases} \tag{4.2}$$

the sample τ-quantile can be found by minimizing $\sum_{i=1}^{N} \rho_\tau(y_i - q)$ with respect to q [51, p. 417]. Figure 4.2 shows the check function for two values of τ. Replacing q with the right hand side of (4.1) leads to the estimates

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau(y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p})) \tag{4.3}$$

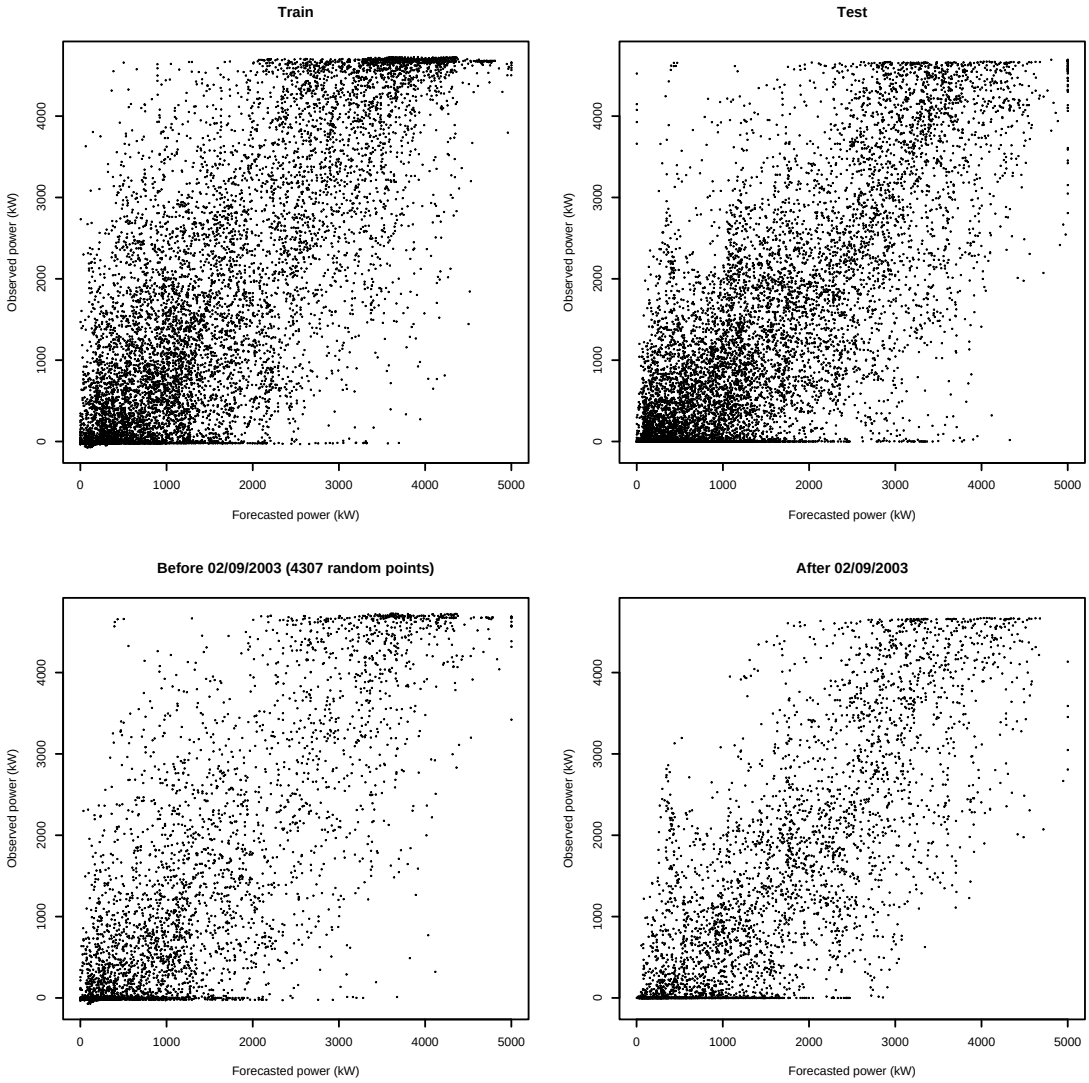

Figure 4.1: *Observed versus forecasted power (horizons 18-36 hours since 06Z) for the training and test data (top row), and before and after a model change in DMI-HIRLAM which is expected to have large influence on the forecasted 10 m wind (bottom row). Note that the length of the training and test periods are both five months. Also, the two plots in the bottom row contain the same number of points; to achieve this a random subset of the points before the model change is selected.*

where $\boldsymbol{\beta}(\tau)$ is a vector containing the unknown coefficients. The estimates can be obtained by used of linear programming techniques [50]. Here we have used the add-on package "quantreg" for "R", see Section 4.1 and A. It is noted that if the check function is replaced with squared loss, i.e. if $\rho_\tau(e) = e^2$, then (4.3) leads to least squares estimates.

4.3.2 Parametric additive quantile models

To simplify the discussion we start by considering models for the mean of a random variable and later we consider models for the quantiles. Generally, when the dependence

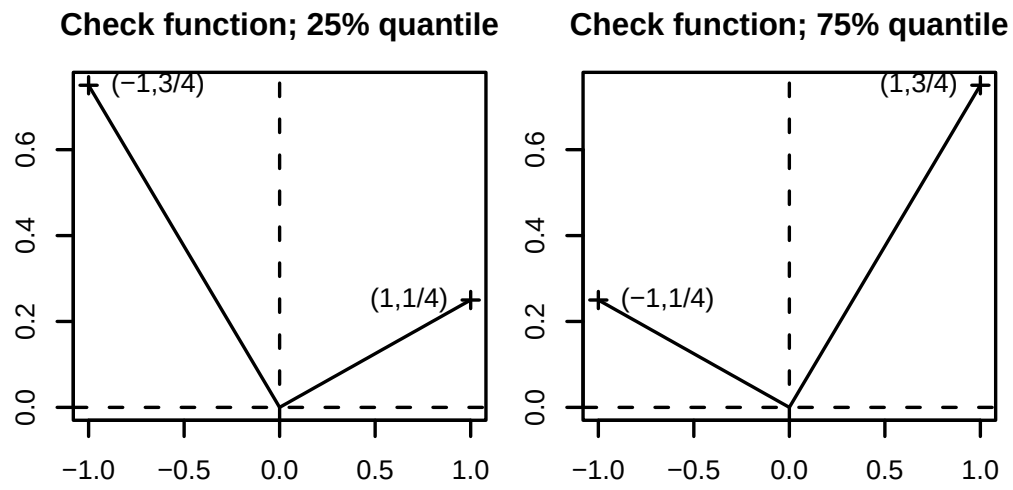

Figure 4.2: *The check function $\rho_\tau(e)$, for $\tau = 0.25$ (left) and for $\tau = 0.75$ (right).*

of y on $x_1, \ldots, x_p$ is not known a very general model is

$$y = g(x_1, \ldots, x_p) + \epsilon \,, \tag{4.4}$$

where g is an unknown function, and ϵ represents iid. errors with mean zero and variance σ^2. In principle it is possible to estimate g, e.g. by use of local regression [20]. However, when investigating many explanatory variables, i.e. more than two or three, the *curse of dimensionality* makes practical application of (4.4) problematic, see [6] and [40, p. 83-4]. To circumvent this problem additive models [40] are used in this chapter. Models of this type can be expressed as

$$y = \alpha + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) + \epsilon \,. \tag{4.5}$$

The constant α and the *functions* $f_\cdot(\cdot)$ can be estimated based on data using non-parametric methods together with backfitting, for details see [40]. However, note that unless the level of functions are restricted the estimates are non-unique, e.g. a constant can be added to one function and subtracted from an other. [40] impose the restriction that each of the function estimates has zero mean over the data.

As described by [40, Sec. 9.3] each of the functions can be approximated by linear combinations of known basis functions of the corresponding explanatory variable, i.e.

$$f_j(x_j) = \sum_{k=1}^{n_k} b_{jk}(x_j)\theta_{jk} \,, \tag{4.6}$$

where $b_{j\cdot}(x_j)$ are the basis functions and $\theta_{j\cdot}$ are unknown coefficients. However, the price paid for the simplicity is that the resulting estimates of the functions generally have larger bias than those based on backfitting and non-parametric methods [40, Sec. 9.3].

It is simple to impose a linear restriction on (4.6), e.g. $f_j(0) = 0$, and derive the resulting $n_k - 1$ basis functions. Likewise, if some of the functions in (4.5) are known to

be periodic this restriction can be imposed on the basis functions. Using e.g. cubic B-spline basis functions [21], the functions in (4.5) have continuous derivatives up to order two. This property should also be imposed when constructing the periodic basis. Plugging (4.6) into (4.5) results in a linear regression model for which the coefficients α and θ_{ik} can be found by use of least squares.

Comparing with (4.1) it is seen that this can be generalized to quantile regression by modelling $Q(\tau)$ as

$$\begin{aligned} Q(\tau) &= \alpha(\tau) + \sum_{j=1}^{p} f_j(x_j;\tau) \\ &= \alpha(\tau) + \sum_{j=1}^{p}\sum_{k=1}^{n_k} b_{jk}(x_j)\theta_{jk}(\tau) \ , \end{aligned} \tag{4.7}$$

with the basis functions constructed under appropriate restrictions on $f_j(\cdot);\ j = 1,\ldots,p$, as outlined above.

In this chapter focus will be on the 25% and 75% quantiles. Since the level of the functions are arbitrary the effect of x_j should be quantified by plotting the sum of the corresponding estimated function and the estimated intercept. To center the plots around zero we subtract the average of the intercepts estimated for the 25% and 75% quantiles. Thus, the effect of x_j is quantified by plotting

$$\hat{f}_j(x_j;\tau) + \hat{\alpha}(\tau) - \frac{\hat{\alpha}(0.25) + \hat{\alpha}(0.75)}{2}$$

for $\tau = 0.25, 0.75$. Otherwise, differences in $\hat{\alpha}(0.25)$ and $\hat{\alpha}(0.75)$, may cause apparent crossings of the 25% and 75% quantiles. The "hat" denotes estimated values.

4.4 Building the quantile model

In this section models for the 25% and 75% quantiles are developed. First a model considering the explanatory variables

pow.fc; forecasted power from WPPT in kW,

horizon; number of hours since 06Z,

ad; forecasted air density from DMI-HIRLAM in g/m^3,

fv; forecasted friction velocity from DMI-HIRLAM in m/s,

wd10m; forecasted wind direction 10m a.g.l. from DMI-HIRLAM in degrees, and

ws10m; forecasted wind speed 10m a.g.l. from DMI-HIRLAM in m/s

is developed. Hereafter, risk indices based on the meteorological forecast variables are considered.

4.4.1 Basic model

Due to the non-linearity of the power curve it is natural to require that the forecasted power production (**pow.fc**) is included in the quantile model of the forecast error. Figure 4.3 shows all pairwise scatter plots of the potential explanatory variables. Due to close relations between some of the variables (**pow.fc**, **fv**, and **ws10m**) it is seen that with the requirement just stated the friction velocity (**fv**) and the 10m wind speed (**ws10m**) can not be included in the model.

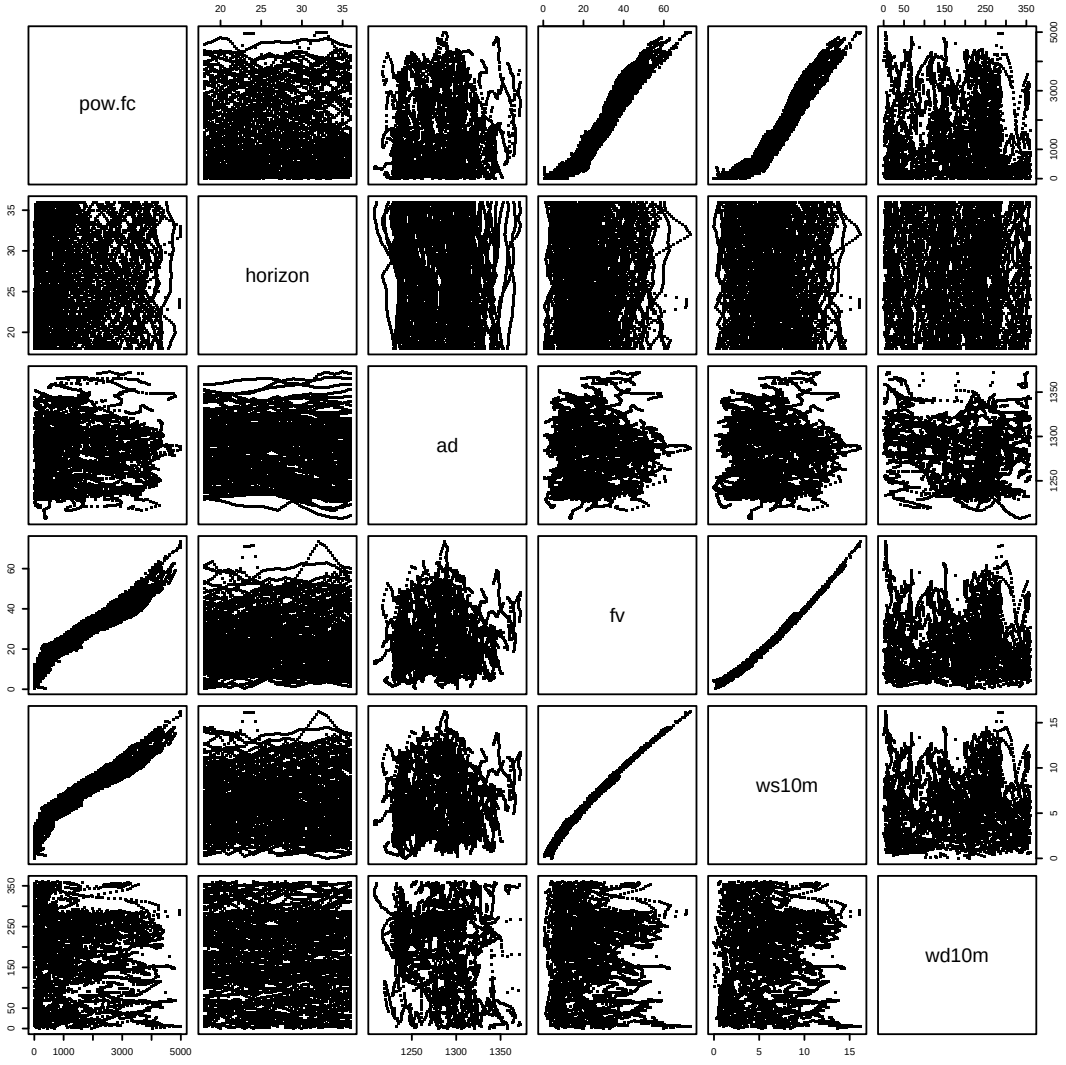

Figure 4.3: *All pairwise scatterplots of the explanatory variables (training data).*

For each of the remaining explanatory variables (**pow.fc**, **horizon**, **ad**, **wd10m**) a spline basis with 10 degrees of freedom is constructed [21]. For the wind direction (**wd10m**) a periodic cubic spline basis with equidistant knots is used. The periodic basis is constructed so that it integrates to zero over the period (360 degrees). For the non-periodic variables natural spline bases without intercepts are used; this implies that the

functions are restricted to be zero at the lower boundary knot. The boundary knots are placed at the limits of the data and the internal knots are placed according to the quantiles of the individual explanatory variables. In this way the model allow for more flexibility where the observations are relatively dense. Note that for prediction it is important to use the same actual knots. Since none of the bases allow for a free intercept this is handled by an intercept in the model. The intercept is expected to vary with the quantile considered.

The resulting model for each of the quantiles (25% and 75%) is depicted in Figure 4.4 which shows the effect of each variable. For each pair of estimates the difference in estimated intercepts is visible for the minimal value of the explanatory variable. It is seen that the effect of horizon is small (almost flat curves) and there is some increased uncertainty for westerly winds. The dependence on air density seems to be minor. Overall the most important explanatory variable is the forecasted power. For the training data crossings of the 25% and 75% quantiles occur in 111 out of 10658 cases.

The estimates describing the dependence on the air density does not seem to have any reasonable interpretation and the differences for low and high densities are supported by very few data points. For this reason it is decided to exclude it from the model. The number of crossings decreases to 46 for the resulting model. The estimates are depicted in Figure 4.5. Since the curves are close when **pow.fc** is zero, it is seen that the intercepts of the 25% and 75% quantile models are very close.

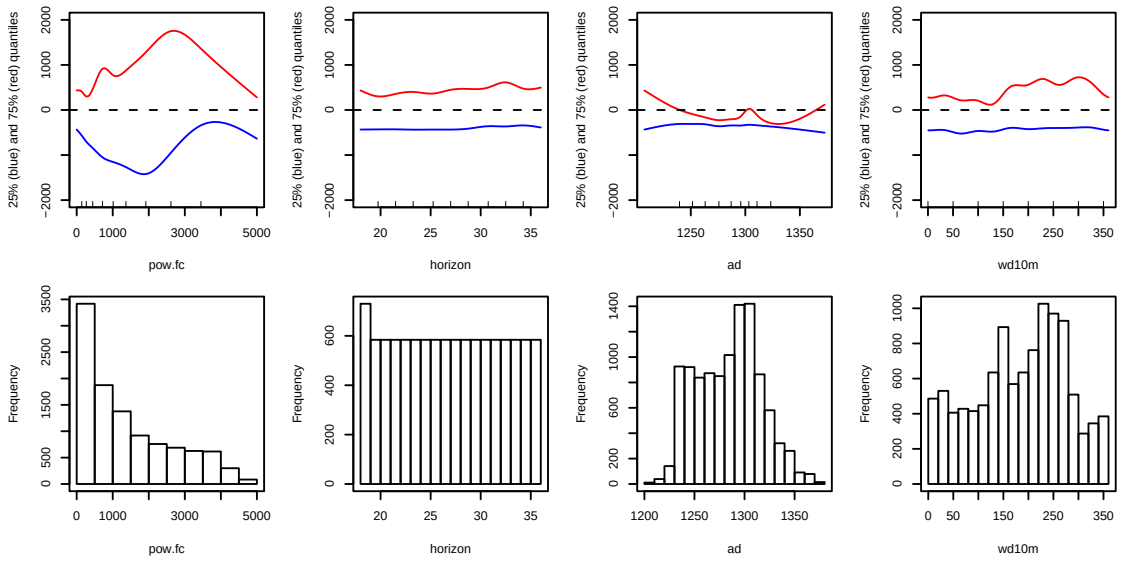

Figure 4.4: *Estimated 25% (blue) and 75% (red) quantiles, together with histograms of the explanatory variables. The rugs on the 1st axis in the top row of plots indicate the placement of knots.*

4.4.2 Risk indices of Meteorological variables

The European Centre for Medium-Range Weather Forecasts (ECMWF), which run the global model supplying boundary conditions for DMI-HIRLAM, perform data assimilation based on 12 hour intervals (00Z and 12Z). For the assessment of the forecast risk the two DMI-HIRLAM forecasts which are based on the two latest global data assimi-

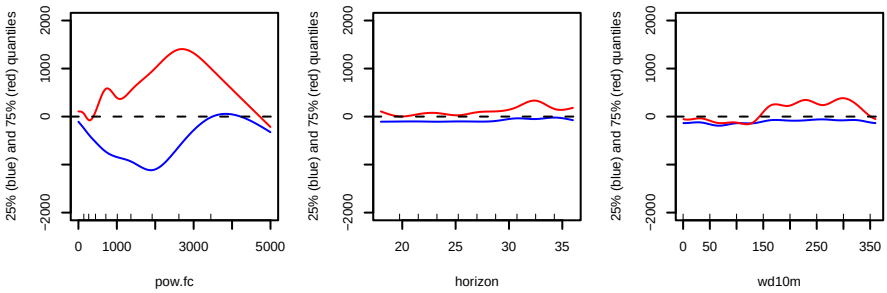

Figure 4.5: *Estimated 25% (blue) and 75% (red) quantiles when excluding the forecasted air density from the model. The rugs on the 1st axis in the top row of plots indicate the placement of knots.*

lations are compared. Since the primary interest is in the 06Z forecast, cf. Section 4.1, this is compared to the preceding 18Z forecast. Since the maximum forecast horizon of DMI-HIRLAM is 48 hours risk indices can only be calculated for the 06Z forecast up to a horizon of 36 hours.

Following [80, 79] the difference in the two forecasts are squared and summed over the entire range of horizons for a particular 06Z forecast. The square root of this number is used as the risk index of each variable, corresponding to each 06Z meteorological forecast. Figure 4.6 shows histograms of the risk indices. It is seen that the risk indices generally only show a few high values; for this reason it is decided only to investigate linear relationships between the quantiles and the risk indices.

When adding the risk indices one at a time to the model shown in Figure 4.5, i.e. without air density, the results shown in Figure 4.7 are obtained. Generally, the risk indices seems to be of minor importance for the quantiles, and it is chosen to use only the one with the clearest signal, i.e. **fv**. Figure 4.8 shows the estimates obtained for this model. It is seen that the effect of the risk index is comparable to the effect of the horizon. For the training data the number of crossings of the 25% and 75% quantiles decreases to 39 for this model.

4.5 Evaluation on test data

The following models are fitted to training data and evaluated on the test data.

Basic: A model using only the forecasted power production as explanatory variables.

Full: The model corresponding to Figure 4.4.

W/o density: The model corresponding to Figure 4.5.

Incl. risk: The model corresponding to Figure 4.8.

The number of crossings on the test data range from 60 to 82 of 11168 cases. In case of crossing of the two quantiles these have been set to their common average. The actual frequencies by which the prediction error is below the 25% quantile or above the 75% quantile in the test data is listed in Table 4.1. A marked difference is seen when splitting

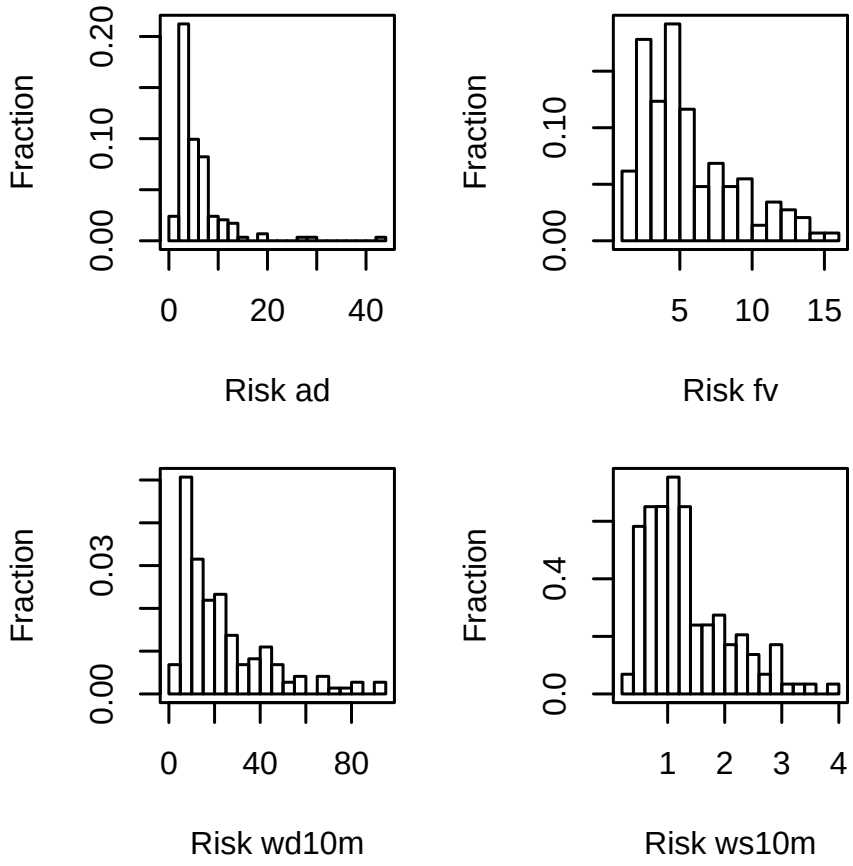

Figure 4.6: *Histograms of risk indices (training data).*

the data according to the date at which an presumably important change was introduced into DMI-HIRLAM. For this reason we focus on the part of the test period up to 2/9-2003.

The model termed "full" seems to result in some problems for the forecasted 75% quantile since this forecast is exceeded in only 10% of the cases. For the model termed "basic" the forecast intervals seems to be symmetric, but too wide. The two remaining models both perform well (52 to 53% change to be between the forecasted quantiles), but the forecast intervals seems to be shifted upwards corresponding to approximately 5%. Random variation may account for some of these differences, but such variations are difficult to quantify due to the inherent and presumably complicated correlation of the data.

Given ensemble quantiles which are correct in a probabilistic sense the quality of these depend on (i) the ability to distinguish between situations with low and high uncertainty and on (ii) the sharpness of the distributions. Here the sharpness is measured as the inter quartile range (IQR), i.e. the difference between the forecasted 75% and 25% quantiles.

Qualitatively (i) is fulfilled if both low and high values of the IQR occur and with respect to (ii) the IQR should be smaller than the IQR obtained from historic production data. These aspects are addressed in Figure 4.9. Results are shown for three models where the basic model is included for reference, although it is not very precise with respect to the

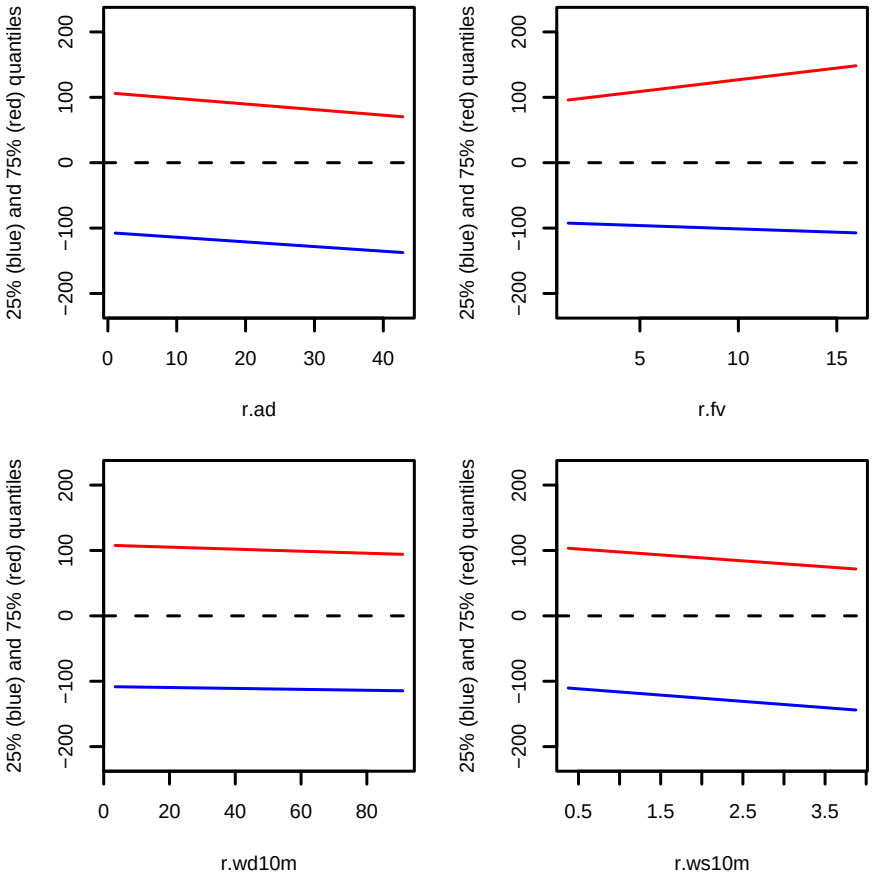

Figure 4.7: *Estimates of the dependence on risk indices when requiring the dependence to be linear and adding the risk indices one at a time to the model depicted in Figure 4.5.*

observed frequencies. It is seen that the basic model differs from the two other models. Also, since the plot for the two other models does not differ markedly, it does not seem very important to include the particular risk index.

The relatively large difference between the 5% and 95% quantiles of the IQR indicates high variability and for probabilistic correct quantiles this can be interpreted as the fulfillment of (i). Furthermore, it is seen that in many situations the ensemble IQR is significantly smaller than the IQR of the historic power productions, i.e. the ensemble forecast is sharp compared to historic data.

4.6 Conclusion and Discussion

We have proposed a method for building models of e.g. the 25% and 75% quantiles of the of forecast errors from existing wind power forecast systems. Such models can be used together with existing systems and has the potential of providing situation-specific information about the uncertainty of a particular forecast.

The quantiles are modelled as a sum of non-linear smooth functions of variables forecasted by the meteorological model or variables derived from such forecasts. The addi-

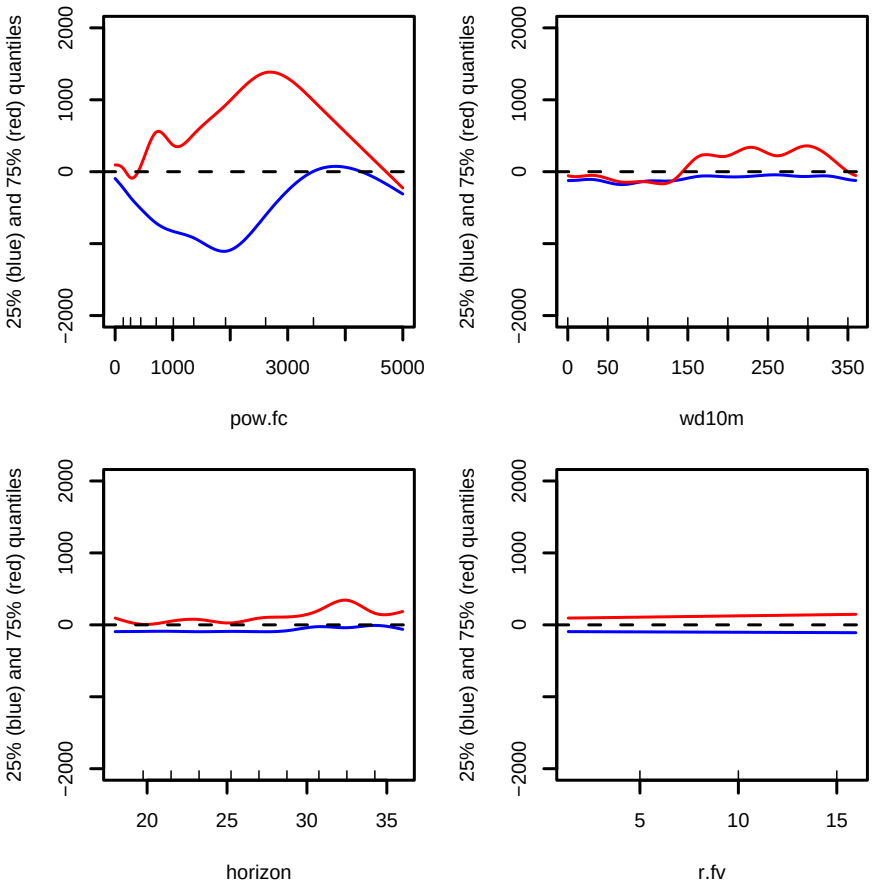

Figure 4.8: *Estimates in the model consisting of the model depicted in Figure 4.5 with the risk index of* ***fv*** *added (bottom, right).*

tive model structure is used since it allows for the inclusion of more explanatory variables than more general non-parametric models. Furthermore, additive models are relatively easy to visualize and interpret. Using spline bases to approximate each of the smooth functions as a linear combination of basis functions only depending on known quantities permits the used of existing linear quantile regression software to fit the models.

The software used is "R", together with the add-on package "quantreg", which can both be freely downloaded from `http://www.r-project.org`. An example R-script is included in Appendix A. "R" could be used to easily extend a given wind power forecast system, and it is even possible to embed "R" into other software products.

With respect to the analysis of the specific data it is noted that the risk indices, which all are inspired by [80, 79], does not seem to have very much influence on the 25% and 75% quantiles. However, it is noted that [80, 79] consider horizons ranging from 0 to 24 hours, whereas we consider horizons ranging from 18 to 36 hours. Note that the horizons mentioned does not take into account the calculation time of HIRLAM. Not surprisingly the most influential variable is the forecasted wind power production. Furthermore, the results show increased uncertainty for westerly winds. Also, the effect of the horizon on the quantiles is minor.

Table 4.1: *Observed frequencies (test data) below forecasted 25% quantile or above the forecasted 75% quantile for the four models considered. Both values for the full test set and for the test set split into two parts based on a presumably important model change in DMI-HIRLAM, cf. Section 4.2.*

	Basic	Full	Without density	Incl. risk
			All test data	
Above	18%	13%	17%	17%
Center	61%	58%	53%	52%
Below	21%	29%	30%	31%
			Test data before 2/9-2003	
Above	19%	10%	19%	19%
Center	62%	64%	53%	52%
Below	19%	26%	28%	29%
			Test data after 2/9-2003	
Above	16%	16%	15%	14%
Center	60%	50%	50%	53%
Below	24%	34%	35%	33%

[10] shows that the optimal quantile to use depend on the actual prices in the market. This will require a range of quantile models to be applied in parallel. Using models with several predictors and spline bases as suggested in this chapter is likely to result in crossing of some of these quantiles. Ideally, the coefficients estimated should be constrained in order to avoid crossings. However, we are not aware of software which can handle this easily. In the situation just outlined it is probably sufficient to use the quantile model indicated by the prices in the market and disregard the fact that this quantile model may cross some other quantile models.

As just outlined quantile regression is characterized by estimating separate models for each quantile. As a consequence crossing of quantiles may occur, indeed in the analysis of the data presented in this chapter a few crossings of the 25% and 75% quantiles occurred. In practice this probably indicates low uncertainty and is therefore of less practical importance. It is however undesirable from a theoretical point of view. One solution would be to start with the median (50% quantile) and find solutions to successive lower and higher quantiles under the restriction that the quantiles does not cross. The restriction should be valid for all possible values of the explanatory variables. Considering the data at hand this could be approximated by considering all observations, i.e. the number of restrictions will be high. We believe that methods based on approximations of the full distribution should also be investigated since this will automatically supply non-crossing quantiles.

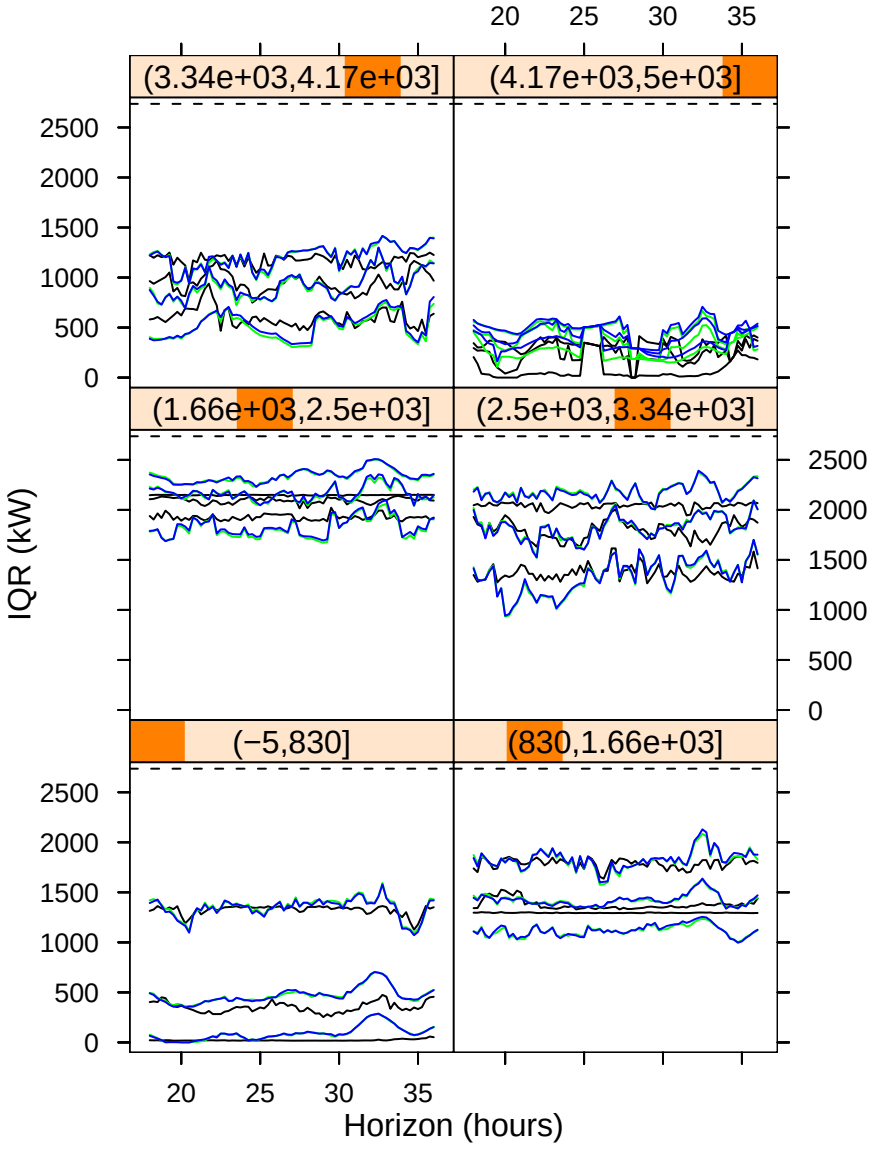

Figure 4.9: *Quantiles (5%, 50%, 95%) of the inter quartile range (IQR). Models: Basic (black), w/o density (green), incl. risk (blue). The grouping variable is the power production (kW) as forecasted by WPPT and split in six intervals of equal length. The horizontal line at 2736 kW indicate the IQR of actual power productions between 1/7-1999 and 1/6-2003.*

Chapter 5

An expert model for the estimation of prediction intervals of wind power

Abstract

In this Chapter is developed and evaluated a generic method appropriate for the estimation of prediction intervals of wind generation. In order to avoid a restrictive assumption on the shape of prediction error distributions, focus is given to an empirical and distribution-free approach. Also, a fuzzy inference model is introduced in order to integrate the expertise on the characteristics of prediction errors for providing conditional interval forecasts. The proposed method can be considered for providing full predictive distributions of wind generation. In parallel, the required properties of probabilistic predictors are given, followed by the description of a non-parametric framework for the verification of wind power probabilistic forecasts in the form of quantiles or intervals. This framework is consequently used for evaluating and analyzing the skill of the proposed approach. This one proves to be reliable and it is shown how its resolution may be enhanced by using the forecaster's expertise. Finally, some guidelines are given for the application of the method to online prediction exercises.

5.1 Introduction

In the present Chapter, our aim is to develop an appropriate method for estimating prediction intervals of wind generation, which can be applied to any state-of-the-art point forecasting method. For that purpose we exploit the characterization of prediction errors carried out in the frame of the Work Package 2 (WP2) of the European project Anemos (cf. Anemos Deliverable Report 2.1 [64]). Since it was shown that these characteristics were shared by all point forecasting approaches (either of the physical or of the statistical type), we use that aspect for developing a generic method. Also, focus is given to the development of a distribution-free approach, suitable for nonstationary, nonlinear and bounded processes. It is explained why such an approach can be applied to any point prediction method. Also, the way it can be straightforwardly applied for the wind generation prediction problem is detailed.

The second part of the Chapter is devoted to the assessment of the quality of the resulting prediction intervals. Quality is related here to a statistical performance follow-

ing Murphy's terminology [68]. Interval forecasts have attracted attention only recently and the assessment of their quality is more complicated than for the case of point predictions. A non-parametric framework for carrying out this performance assessment is introduced. Then, an evaluation of the derived method quality is given by applying the proposed framework to various case-studies and on several state-of-the-art point prediction methods. These case-studies are the Klim and Tunø Knob wind farms, that were already considered in WP2. Also, the three different methods we use for point forecasting were evaluated in the frame of WP2. These prediction approaches are denoted by M1, M2 and M3. In complement, we highlight the influence of the parameters of the introduced interval forecasting method on some particular aspects of the quality of prediction intervals. This allows us to derive guidelines for the application of the developed approach to online forecasting exercises.

5.2 Different types of statistical intervals

Often, it is needed to draw conclusions on the characteristics of a process from a limited amount of available knowledge. Statistics are usually calculated from limited samples and may prove to be uncertain. Perhaps the most illustrative example is that of public opinion polls, for which panels composed by few hundreds or thousands people are used to tell what is the average opinion of millions of people in a country. Since this population sampling induces uncertainty, calculated statistics are therefore associated with estimates of their accuracy, in the form of intervals. Depending on the type of decision to make from a given statistic, several types of related intervals may be defined. For an introduction to these different types of statistical intervals, we refer to [37].

Our concern here is about the accuracy of point forecasts. Two types of intervals appear to be relevant for that purpose: *confidence intervals* and *prediction intervals*. There is a fundamental difference between these two. Given a sample population $\{p_t\}_{t=1,\dots,T}$, a confidence interval is meant for giving a measure of confidence on the estimate $\hat{\theta}(\{p_t\}_{t=1,\dots,T})$ of a parameter θ for the whole population, whereas a prediction interval is meant for giving the range of values within which the next randomly selected individual p_t ($t > T$) from that population may lie, with a certain degree of confidence.

In order to describe how this can be translated to the forecasting problem, let us consider the case of a statistical model g designed for 1-step ahead prediction. The parameters $\mathbf{w}$ of that statistical model are estimated over a training set consisting of N_L pairs $\{\mathbf{y}_t, p_t\}_{t=1,\dots,N_L}$, where $\mathbf{y}_t$ is a vector containing past values of p (up to time $t-1$) plus eventually past values and forecasts of explanatory variables, and p_t is the observation at time t. In a general manner, $\mathbf{y}_t$ includes a number of past values p_{t-i} ($i = 1, \dots, l$) of the variable of interest, plus past values of some explanatory variables $\mathbf{x}_{t-i}$ ($i = 1, \dots, m$), and eventually forecast values of these explanatory variables $\hat{\mathbf{x}}_{t/t-1}$. Writes

$$\mathbf{y}_t = (p_{t-1}, p_{t-2}, \dots, p_{t-l}, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \dots, \mathbf{x}_{t-m}, \hat{\mathbf{x}}_{t/t-1}). \tag{5.1}$$

These data pairs are assumed to be generated according to the following process:

$$p_t = g(\mathbf{y}_t, \mathbf{w}) + e_t, \tag{5.2}$$

where $\mathbf{w}$ are the parameters of the chosen model g and $\{e_t\}$ is a zero mean random variable. For $t > N_L$ the trained model $g(\mathbf{y}_t, \hat{\mathbf{w}})$ (with $\hat{\mathbf{w}}$ estimated by considering a quadratic loss function) will then produce at time t a forecast $\hat{p}_{t+1/t}$ that is an estimate of the mean $\bar{p}_{t+1}$ of the target distribution F^p_{t+1} at time $t+1$, given $\mathbf{y}_{t+1}$[1]. The uncertainty in the estimate of the mean of the target distribution partly comes from the fact that one uses a finite sample for training the model, which consists in an incomplete knowledge of the true process. In addition, observations may integrate a noise component coming from data acquisition devices. This uncertainty is also due to the choice of the model that may not reflect the true behavior of the process and to the way the model parameters $\mathbf{w}$ are estimated. A confidence interval associated to $\hat{p}_{t+1/t}$ is hence a measure of the confidence in the estimation of the mean $\bar{p}_{t+1}$ of the target distribution. Since $\hat{p}_{t+1/t}$ is not an estimate of the true outcome p_{t+1}, this interval does not give the confidence in the estimation of the true effect p_{t+1}.

Alternatively, a prediction interval associated to a point forecast is a measure of the accuracy of that point forecast with respect to the true outcome p_{t+1}, by giving a range of potential values. A prediction interval necessarily encloses the corresponding confidence interval [41]. Figure 5.1 is an illustrative example of the difference between confidence and prediction intervals. The solid curve represents a probability distribution of expected wind generation at time $t+1$. The two dashed vertical lines correspond respectively to the mean $\bar{p}_{t+1}$ of that distribution (bold) and to a wind power point forecast $\hat{p}_{t+1/t}$, which is an estimate at time t of that mean. The dark shaded area stands for the confidence interval associated to $\hat{p}_{t+1/t}$, while the light shaded area is for the interval forecast. The solid vertical line gives the observed power value at time $t+1$.

Although we have taken the example of a statistical model designed for 1-step ahead predictions, this reasoning can be extended to the case of other types of models (i.e. multi-step ahead models and physical models): they all aim at estimating a particular point of the target distribution, which is its mean in most of the cases. Then, a confidence interval will always correspond to the confidence in the estimate of the expected outcome, whereas a prediction interval associated to a point forecast will give the accuracy of that estimate with respect to the true outcome. Because we are mostly interested in that second type of uncertainty assessment we will turn our attention to prediction intervals from now on.

Formally, a *prediction interval* $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$, alternatively called interval forecast, estimated at time t for lead time $t+k$, is a range of values within which the true effect p_{t+k} is expected to lie with a certain probability $(1-\alpha)$, denoting its *nominal coverage rate*:

$$\mathrm{P}\left(p_{t+k} \in \hat{\mathrm{I}}^{(\alpha)}_{t+k/t}\right) = \mathrm{P}\left(p_{t+k} \in [\hat{L}^{(\alpha)}_{t+k/t}, \hat{U}^{(\alpha)}_{t+k/t}]\right) = 1-\alpha. \tag{5.3}$$

[1] See [18] for a proof in the specific case of univariate processes. It is exlained how this proof may be straightforwardly extended to the case of multivariate processes such as the one we consider here.

An interval forecast is then specified by its respectively lower and upper bounds $\hat{L}^{(\alpha)}_{t+k/t}$ and $\hat{U}^{(\alpha)}_{t+k/t}$. Note that we will prefer the term 'nominal coverage rate' (or alternatively 'degree of confidence') instead of the widely used 'confidence level' term when referring to the probability associated to interval forecasts, so that the reader does not confuse them with the more classical confidence intervals.

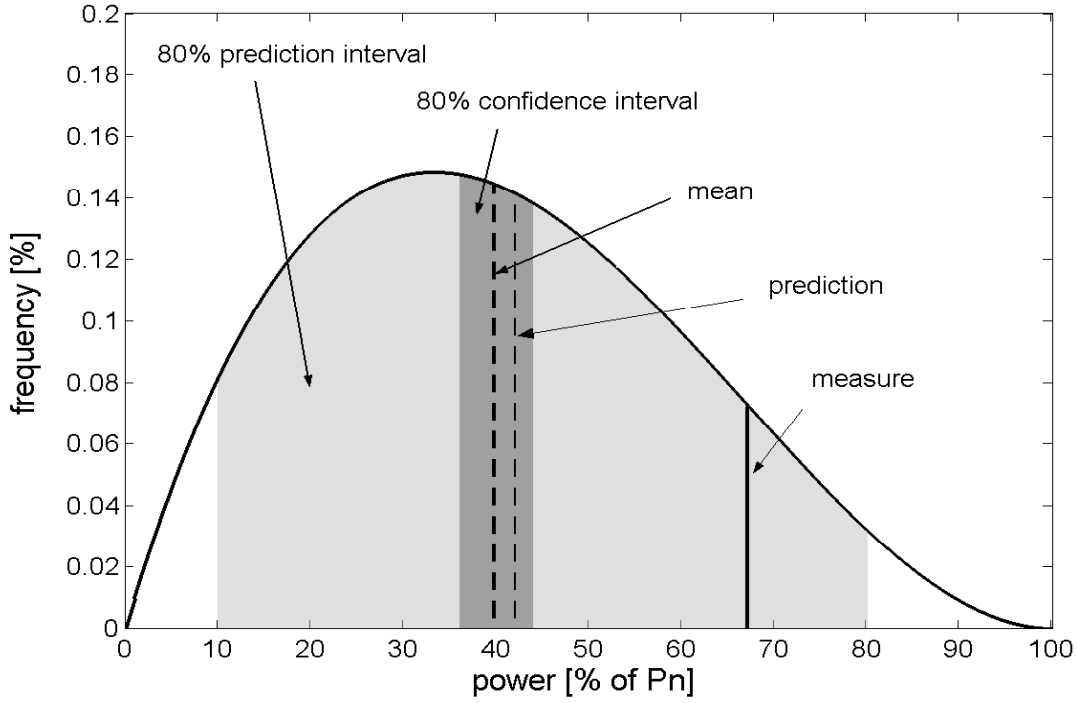

Figure 5.1: *Illustrative example of the difference between confidence (light shaded area) and prediction (dark shaded area) intervals. The confidence interval is a measure of the confidence in our estimate $\hat{p}_{t+1/t}$ of the expectation $\bar{p}_{t+1}$, whereas the prediction interval is related to the accuracy of the point forecast $\hat{p}_{t+1/t}$ with respect to the true effect p_{t+1}.*

Most of the times prediction intervals are central prediction intervals: there is the same probability $(\alpha/2)$ for a non-covered outcome to be above or below the interval bounds. Then, these bounds correspond to the quantiles[2] with proportion $(\alpha/2)$ and $(1-\alpha/2)$ of the *predictive distribution* $\hat{F}^{p}_{t+k/t}$ of future events:

$$\hat{L}^{(\alpha)}_{t+k/t} = \hat{r}^{(\alpha/2)}_{t+k/t}, \quad \mathrm{P}\left(p_{t+k} < \hat{L}^{(\alpha)}_{t+k/t}\right) = \alpha/2, \tag{5.4}$$

$$\hat{U}^{(\alpha)}_{t+k/t} = \hat{r}^{(1-\alpha/2)}_{t+k/t}, \quad \mathrm{P}\left(p_{t+k} < \hat{U}^{(\alpha)}_{t+k/t}\right) = 1-\alpha/2. \tag{5.5}$$

Central prediction intervals are hence centered on the median of the predictive distribution $\hat{F}^{p}_{t+k/t}$.

[2]The quantile $r^{(\alpha)}$ with proportion α of the distribution F^X of a random variable X is defined as the value x such that $\mathrm{P}(X \leq x) = \alpha$.

Traditionally, emphasis is given in the literature to the computation of prediction intervals for a Normal distribution, or more generally for a symmetric target distribution [13, 22, 24, 37, 41]. Thus, estimated prediction intervals are centered on the point prediction itself and give the equally probable upward and downward margins in which the future outcome may lie. Due to symmetry, the mean and median of these target distributions are equal. Moreover, the upper and lower sides of the intervals have the same size. Therefore, whatever the nominal coverage rate, the point forecast is included in the interval forecast it is associated to. For a nonlinear and bounded process such as wind generation, probability distributions of future power output exhibit some skewness. For these asymmetric distributions, the median may significantly differ from the mean, and thus central prediction intervals (for rather low nominal coverage rate) may not even cover the point forecast value. This is why interval forecasts can be alternatively constructed in the form of intervals $\hat{\mathrm{I}}\mathrm{c}^{(\alpha)}$ centered on the point forecast itself

$$\hat{\mathrm{I}}\mathrm{c}^{(\alpha)}(\hat{p}_{t+k/t}) \;=\; [\hat{L}c^{(\alpha)}_{t+k/t}, \hat{U}c^{(\alpha)}_{t+k/t}], \tag{5.6}$$

as equally probable positive and negative margins in which the actual outcome may lie, for a given nominal coverage rate $(1-\alpha)$:

$$\mathrm{P}\left(p_{t+k} \in [\hat{L}c^{(\alpha)}_{t+k/t}, \hat{p}_{t+k/t}]\right) \;=\; \mathrm{P}\left(p_{t+k} \in [\hat{p}_{t+k/t}, \hat{U}c^{(\alpha)}_{t+k/t}]\right) \;=\; (1-\alpha)/2. \tag{5.7}$$

Such a type of intervals will be referred to as prediction-centered interval forecasts. They consist in separately modeling two different probability distributions, which are the ones for the respectively positive and negative errors. Then, one notes that even if $\hat{L}c^{(\alpha)}_{t+k/t}$ and $\hat{U}c^{(\alpha)}_{t+k/t}$ are quantiles of the whole predictive distribution, we do not know the proportions they correspond to. Since we aim at directing our work towards a probabilistic view of wind power forecasting, our preference goes to central prediction intervals, since they model the target distribution F^p_{t+k} as a whole. Consequently, by specifying a nominal coverage rate $(1-\alpha)$, we will then determine the quantiles with proportions $(\alpha/2)$ and $(1-\alpha/2)$ of $\hat{F}^p_{t+k/t}$.

Finally, as for point forecasts, prediction intervals issued at time t are produced from the information set Φ_t that gathers the available information up to that time. Therefore, even if we use the notation $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$ in the following of the Chapter, it actually stands for $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}(\Phi_t)$.

5.3 Basic parametric approaches for prediction interval estimation

An approach is said to be *parametric* if there is an underlying assumption on the distribution one tries to model. Inversely, a *non-parametric* (or *distribution-free*) approach does not rely on such an assumption.

The simplest parametric approach for estimating prediction intervals is the method proposed by Box and Jenkins [8]. It follows the assumption that for a model (such as the

multivariate one given by Equation (5.2)) the $\{e_t\}$ sequence is independent and identically distributed Gaussian with zero mean and variance $\sigma_e^2 < \infty$, $e_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2)$. By using an estimate $\hat{\sigma}_e^2$ of the variance, the one-step ahead interval forecasts with nominal coverage rate $(1-\alpha)$ are such that:

$$\hat{\mathrm{I}}_{t+1/t}^{(\alpha)} = \left[\hat{p}_{t+1/t} - z_{\alpha/2}.\hat{\sigma}_e^2,\, \hat{p}_{t+1/t} + z_{1-\alpha/2}.\hat{\sigma}_e^2\right], \tag{5.8}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the quantiles with proportion $(\alpha/2)$ an $(1-\alpha/2)$ of the standard Normal distribution $\mathcal{N}(0,1)$. Then, k-step ahead interval forecasts can be produced similarly, by considering estimates of the variance $\hat{\sigma}_{e,k}^2$ of the random shock for k-step ahead point predictions

$$\hat{\mathrm{I}}_{t+k/t}^{(\alpha)} = \left[\hat{p}_{t+k/t} - z_{\alpha/2}.\hat{\sigma}_{e,k}^2,\, \hat{p}_{t+k/t} + z_{1-\alpha/2}.\hat{\sigma}_{e,k}^2\right]. \tag{5.9}$$

By assuming that errors in consecutive step-ahead forecasts are mutually independent and distributed Gaussian with zero mean and constant variance, Makridakis et al. [63] proposed to compute the prediction intervals of Equation (5.9) with an 'approximate' formula for $\hat{\sigma}_{e,k}^2$, which states that

$$\hat{\sigma}_{e,k}^2 = k.\hat{\sigma}_e^2. \tag{5.10}$$

It has been shown by Koehler [48] that there was no theoretical justification for Equation (5.10), and that the assumptions mentioned above about the error process could only be true for a random walk model. Therefore, using such a simple approximation of the k-step variance would yield very inadequate results.

Instead of relying on approximate formulae for estimating the k-step ahead variance, another possibility is to use the historical performance of the predictor:

$$\hat{\sigma}_{e,k}^2 = \mathrm{SDE}^2(k), \tag{5.11}$$

where $\mathrm{SDE}(k)$ is the standard deviation of the k-step ahead forecasting errors over a given evaluation period, as defined in [62]. Alternatively, one may consider the use of time-adaptive statistics for estimating recent SDEs of the prediction method.

Intervals estimated from Equation (5.9) are symmetric around the point prediction. Even if the Gaussian assumption does not hold, the Box-Jenkins method is often followed in practice. When considering nonlinear (and chaotic) time-series such a basic estimation of prediction interval bounds will lead to poor results [52]. This has recently been illustrated for the specific case of wind power forecasting [78].

We know that the nonlinearity aspect is due to the energy conversion process. When thoroughly studying conditional distributions of wind speed prediction errors (given predicted wind speed), Lange [56] noticed that they could be modeled with Gaussian distributions whose standard deviations equal the standard deviation of unconditional error distributions of wind speed forecasts. In parallel, he proposed a model based on the local derivative of the wind park power curve for describing the way wind speed intervals

would be mapped to power intervals. He used that model for estimating the standard deviation of conditional distributions of power prediction errors given predicted power output. Then, a Gaussian assumption was considered for calculating the $(1-\alpha)$ prediction interval of wind generation [55]. The first shortcoming of this approach is that power prediction errors are assumed to be distributed Gaussian. This could be easily overcome by estimating $(1-\alpha)$ intervals on the error distributions of wind speed and by passing these intervals through the wind park power curve for obtaining non-Gaussian intervals of wind power. This idea has already been proposed by Brown et al. [12]. The second shortcoming is that the method does not account for the modeling error itself, owing to the spatial refinement of the NWPs or to the model used for the power curve. Also, such a method is limited for application to physical methods only since it requires an explicit power curve. Finally, standard deviations of wind speed error distributions are not easy to obtain since wind forecasts and related measurements are often not available at the level of a wind farm. They may be provided as a guess by meteorological offices based on their expertise, but it is unlikely that resulting prediction intervals would be accurate.

The nonlinear and bounded nature of the wind generation process is taken into account by the method proposed by Luig et al. [61], which is based on modeling predictive distributions of power output using β-distributions. Such distributions are bounded between 0 and 1 (like is normalized power production) and their shape is controlled by two parameters α and β. These two parameters are a function of the mean and the variance of the distribution. Luig et al. proposed to set the mean of predictive β-distributions equal to point forecast values given by a power prediction approach. In parallel, the variance of these distribution is determined from a study of the historical performance of the considered prediction method. Different estimates of the variance are considered depending on the range of predicted power (i.e. four ranges in this case). Central prediction intervals are provided consequently by quoting quantiles of estimated predictive distributions. This approach is expected to offer a significant improvement against intervals produced from a Gaussian assumption. However, considering only certain variance estimates for some ranges of predicted power values does not reflect the continuous variability of the power prediction uncertainty, as described in [64]. Moreover, the choice of β-distributions is also a restrictive assumption on the predictive distributions of wind generation.

5.4 Development of a distribution-free approach appropriate for non-linear and bounded processes

When it is not possible to use theoretical formulae, and in the case for which the hypothesis that prediction errors follow a known distribution appears to be a weak assumption, an alternative solution is to develop *non-parametric* approaches for the estimation of predictive distributions or interval forecasts [15]. More generally, distribution-free approaches are appealing since they are not related to any assumption concerning the error-generating process, i.e. to a particular model. Therefore, they are suitable for estimating the uncertainty of different types of forecasting methods either of the statistical or of the physical types. This is also valid if forecasts are the results of some combination

procedure [89].

Quantile regression is a family of non-parametric methods that aim at estimating predictive quantiles for a given proportion. It has recently been considered for producing interval forecasts of wind generation [10, 73] in two different manners. Nielsen et al. [73] proposed a quantile regression method that uses as input point forecasts produced by a state-of-the-art forecasting method, plus some other explanatory variables e.g. the wind speed and direction forecasts that were previously utilized by the point prediction tool. That method can hence be considered for application to already installed point prediction methods in order to associate point forecasts with an estimation of their accuracy. Alternatively, Bremnes [10] developed a slightly different approach, which is based on linear quantile regression with only NWPs as input. This approach has the advantage that one then avoids the point prediction step for producing interval forecasts of wind generation. An important shortcoming of quantile regression approaches is that a specific model needs to be set-up and trained for every quantile of the predictive distribution to be estimated. Therefore, one already has to consider two different models for estimating a single prediction interval. And, for having an adequate estimation of complete predictive distributions, say by forecasting quantiles for every 0.05 proportion, this would lead to 19 models (!). Nielsen et al. [73] pointed out that since models are independently trained, they may result in inconsistent results in certain situations e.g. crossing quantiles. This is not desirable from both a theoretical and practical point of view. Moreover, these models are site-dependent: for each new wind farm they are applied to, a dataset must be collected in order to estimate their parameters through a training process.

In the present work, our aim is to propose not only a non-parametric approach, but also an approach that can be utilized for easily estimating multiple power prediction intervals (and thus several quantiles) at once. Consequently, the target method has to directly construct the predictive distribution of wind generation at once — this was one of the conclusions by Nielsen et al. [73] for avoiding the crossing-quantile problem. This is possible if one considers *empirical* approaches such as the one developed in the following. In a first stage, we introduce the main assumptions related to empirical approaches for prediction interval estimation, as well as the underlying methodology. Then, our contribution is to propose an upgrade of the empirical methods introduced in the literature, which is appropriate for non-linear and bounded processes such as wind generation. Paragraph 5.4.2 describes the classification of forecast conditions related to different characteristics of prediction error distributions. The fuzzy inference model developed in Paragraph 5.4.3 provides conditional distributions of prediction errors as a function of forecast conditions, in the form of combined probability distributions. By dressing point predictions with the estimated conditional distributions of prediction errors, one obtains predictive distributions of wind generations. Finally, two approaches for the combination of empirical distribution functions are given in Paragraph 5.4.4.

5.4.1 Hypothesis and development of empirical methods for prediction interval estimation

In general, as 'empirical' are characterized methods that origin from knowledge or experience. Here, the empirical nature of the prediction interval estimation method stands for the fact that interval forecasts are produced from the witnessed behavior of the point forecasting method it is applied to. The behavior of the point prediction approach is characterized by its recent performance.

The development and application of empirical-type approaches for prediction interval estimation can be traced back to the works by Williams and Goodman [92]. The authors fitted a regressive model for producing 18-step ahead forecasts of the number of phone lines in service over a dataset consisting of 169 data points, and envisaged to associate them with an estimation of their accuracy. Therefore, they estimated prediction intervals with the method described hereafter, by assuming that future forecast errors would be distributed in the same way than the recent ones. They noticed that prediction errors were not Normally distributed — they actually seemed to follow a Γ-distribution. And, despite the rather limited dataset, they showed that this basic empirical approach was much more efficient than usual Box-Jenkins methods for estimating prediction intervals, for various degrees of confidence. The Williams-Goodman method has been applied (with minor changes) on some other forecasting exercises for which prediction errors proved to be not-Normally distributed [45]. More particularly, Alves da Silva and Moulin [1] used a similar method for estimating prediction intervals associated to point forecasts produced with a neural-network-based method, for the short-term load forecasting problem. The authors compared the empirical interval forecasts with two other approaches, namely 'error output' (which is based on a second neural-network model trained for estimating the prediction error of the first neural network) and 'multilinear regression' (which is based on a regressive model with variables the output of the hidden layer neurons and coefficients the weights of the output neuron for estimating the prediction error variance — intervals are consequently computed following the Box-Jenkins method). Conclusions of the study were in favor of the use of the empirical approach.

The first step before computing prediction intervals is to collect the prediction errors the method made in the past. The intervals that are going to be computed will rely on the most recent information on the method's performance. For that purpose, a window in the past (a certain number of hours) is defined and used as a sliding window for storing the errors. The size n of this window determines the size of the samples of errors. At time t, a separate sample $S_{t,k}$ is defined for each prediction horizon k (i.e. for 1-hour ahead, 2-hour ahead, and so on) since we have shown that prediction uncertainty significantly varies with the look-ahead time. The collected errors are the most recent ones at a given time: when the actual measured wind power is known, that value is compared with all the past predictions made for that time. Using the most recent information on a given method performance for estimating future uncertainty is motivated by the non-stationary aspect of wind power prediction errors. Write $\Omega_{t,k}$ the set of prediction errors associated to k-step ahead point predictions up to time t:

$$\Omega_{t,k} = \{\epsilon_{t-i+k/t-i},\, i \in \mathbb{N},\, i \geq k\}, \tag{5.12}$$

where $\epsilon_{t-i+k/t-i}$ is the normalized prediction error related to the point forecast $\hat{p}_{t-i+k/t-i}$. Since the wind generation process is bounded, we will hereafter only deal with normalized errors and predicted values (both normalized by P_n). Straightforwardly, by renumbering the elements of $\Omega_{t,k}$, an error sample $S_{t,k}$ containing the last n k-step ahead point prediction errors at time t consists in

$$S_{t,k} = \{\epsilon_i \in \Omega_{t,k},\, i = 1, \ldots, n\}. \tag{5.13}$$

The *empirical distribution function* $\hat{F}^{\epsilon}_{t,k}$ of errors, at time t and for horizon k, is defined as the discrete distribution that puts probability $1/n$ on each element of $S_{t,k}$. It can be shown that $\hat{F}^{\epsilon}_{t,k}$ is the non-parametric maximum likelihood estimate of the true distribution function of errors $F^{\epsilon}_{t,k}$ (see [26], p. 310). Consequently, any parameter $\hat{\theta}(\hat{F}^{\epsilon}_{t,k})$ estimated from $\hat{F}^{\epsilon}_{t,k}$ is the non-parametric maximum likelihood estimate of the parameter $\theta(F^{\epsilon}_{t,k})$. For practical use, we introduce the cumulative distribution function $\hat{G}^{\epsilon}_{t,k}(\epsilon)$, which gives the fraction of errors less than or equal to ϵ

$$\hat{G}^{\epsilon}_{t,k}(\epsilon) = \frac{1}{n}\, \#\{\epsilon_i \in S_{t,k} \,|\, \epsilon_i \leq \epsilon\}. \tag{5.14}$$

The underlying assumption of the empirical approach is that future uncertainty can be expressed from the recently witnessed behavior of the point prediction method. This means that we consider here that the empirical distribution function of errors $\hat{F}^{\epsilon}_{t,k}$ can be seen as an estimate of the distribution of errors associated to the point forecast $\hat{p}_{t+k/t}$. Therefore, an empirical predictive distribution $\hat{F}^{p}_{t+k/t}$ of wind power output at lead time $t+k$ can be constructed as following:

$$\hat{F}^{p}_{t+k/t} \rightarrow \{\hat{p}_{t+k/t} + \epsilon_i,\, \epsilon_i \in S_{t,k}\}, \tag{5.15}$$

with an equal probability $1/n$ associated to each element of $\hat{F}^{p}_{t+k/t}$.

Since the bounds of the central prediction interval $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$ with nominal coverage rate $(1-\alpha)$ are defined as the quantiles with proportion $(\alpha/2)$ and $(1-\alpha/2)$ of the predictive distribution $\hat{F}^{p}_{t+k/t}$, they are given by:

$$\hat{L}^{(\alpha)}_{t+k/t} = \hat{p}_{t+k/t} + {\hat{G}^{\epsilon}_{t,k}}^{-1}(\alpha/2), \tag{5.16}$$

$$\hat{U}^{(\alpha)}_{t+k/t} = \hat{p}_{t+k/t} + {\hat{G}^{\epsilon}_{t,k}}^{-1}(1-\alpha/2). \tag{5.17}$$

Such a construction of the predictive distribution $\hat{F}^{p}_{t+k/t}$ of wind generation from recent performance implicitly assumes the *representativeness* of the sample data. Actually, this hypothesis cannot be completely exact and then the prediction intervals may only provide a lower bound on the real forecast uncertainty. Note that parametric interval estimation methods described in Section 5.3 also assume that near-future uncertainty will be like the historical one, since estimates of the error variances are based on past performance of the considered prediction method. Secondly, it is implicitly assumed that the sample is a *random sample*, that we do not apply any selection procedure that will then introduce a bias in the uncertainty estimation. This assumption is also not completely

respected for the case of k-step wind power forecasting since consecutive prediction errors may be correlated[3] [70]. However, we will see that breaking this assumption will not have a significant influence on the performance of prediction intervals of wind generation.

5.4.2 Classification of forecast conditions

When predicting nonlinear processes, it is of common knowledge that the shape of the prediction error distributions evolves as a function of the value of the variable of interest [7, 43]. For the specific case of wind power forecasting, there may also be other variables that have an impact on the characteristics of forecast error distributions. We will refer to these variables as *influential variables*. They obviously include predicted power but they may also include forecast wind speed and direction, and eventually some other explanatory variables that are expected to have an influence on the characteristics of prediction error distributions. However, even by applying an empirical approach such as the one presented in the previous Paragraph, prediction intervals will be estimated in the same way whatever the level of influential variables: they actually are *unconditional* interval forecasts. It is unlikely that samples of prediction errors would be representative of the current — and thus conditional — uncertainty. An illustrative example would be the case where collected errors correspond to situations for which the level of predicted power was low and where the current power prediction is in the medium power range. It is hence necessary to propose a more dynamic approach that would be appropriate for estimating *conditional* prediction intervals. Our proposal is then to enhance the empirical method initially described by Williams and Goodman [92] for giving an assessment of the prediction uncertainty related to current forecast conditions. The present Paragraph concentrates on the classification of these forecast conditions.

We define as a *forecast condition* $c_{t,k}$ at time t and horizon k the association of a set of values of the considered influential variables. Denote by $v^l_{t,k}$ the l^{th} influential variable (say that we consider L different variables, hence $l = 1, \ldots, L$) related to the point prediction $\hat{p}_{t+k/t}$. We make the assumption that all the influential variables are bounded[4] and can thus be normalized. Consequently, we have

$$v^l_{t,k} \in V_l = [0, 1] \quad \forall l, t, k. \tag{5.18}$$

Prediction errors are also normalized and bounded, though they lie in the range $[-1, 1]$.

What we referred to as a forecast condition at time t for lead time $t + k$ is uniquely defined by the association of the values of each of the L influential variables:

$$c_{t,k} = \{v^1_{t,k}, v^2_{t,k}, \ldots, v^L_{t,k}\}, \quad c_{t,k} \in \mathcal{C} = V_1 \times V_2 \times \ldots \times V_L, \tag{5.19}$$

[3] Actually, there exists a correlation between prediction errors for successive look-ahead times i.e. between $e_{t+k/t}$ and $e_{t+k+i/t}$, $i > 0$, as well as between predictions for the same look-ahead time but issued at consecutive time origins i.e. between $e_{t+k/t}$ and $e_{t+k+i/t+i}$, $i > 0$. Here, our concern is mainly about the first type of correlation, since interval forecasts are estimated independently for each prediction horizon.

[4] This assumption about the bounded nature of influential variables appears reasonable: the range of physically possible values for both measured or forecast variables obviously have a lower and an upper bound. If outside of that range, these values can be deemed as suspicious or even as outliers.

where $\mathcal{C}$ is the set of possible forecast conditions at any time t and look-ahead time k.

Then, we map $\mathcal{C}$ with a finite number of subsets to which are associated different kinds of characteristics of prediction error distributions. For that purpose, consider J_l ranges of possible values for each of the influential variables v^l $(l = 1, \ldots, L)$. Consequently, we define as $V_l^{j_l}$ the subset of V_l that contains the variable values in the ${j_l}^{th}$ range. By construction, V_l is the union of all of its subsets

$$V_l = V_l^1 \cup V_l^2 \cup \ldots \cup V_l^{J_l}, \quad \forall l, \tag{5.20}$$

such that none of these subsets are overlapping

$$V_l^i \cap V_l^j, = \emptyset, \quad \forall l, i, j, \, i \neq j. \tag{5.21}$$

Now that the sets of possible values for the various influential variables are split into subsets accounting for different characteristics of prediction error distributions, $\mathcal{C}$ can also be split into all the possible associations of the subsets for the various influential variables. Write

$$\mathcal{C}(\{(l, j_l)\}) = \mathcal{C}((1, j_1), \ldots, (L, j_L))) = V_1^{j_1} \times V_2^{j_2} \times \ldots \times V_L^{j_L}, \quad \forall j_l, \tag{5.22}$$

these subsets corresponding to the ${j_l}^{th}$ range of values for each of the L different influential variables. This hence yields N_s subsets, where

$$N_s = \prod_{l=1}^{L} J_l. \tag{5.23}$$

If for instance one considers two influential variables (say forecast wind power and forecast wind direction) for which sets of possible predicted values are split into two subsets, then $\mathcal{C}((1,1),(2,2))$ corresponds to the subset of forecast conditions for which predicted wind power lies in its first subset and predicted wind direction in its second subset. Again, by construction, $\mathcal{C}$ is the union of all of its subsets, such that none of them are overlapping. Note that this classification of the forecast conditions with different related characteristics of prediction error distributions can only be the result of a thorough analysis of the error-generating process. Analyses of forecasting errors are often very informative and allow the analyst to gain expertise on the prediction problem.

Since our aim is to associate specific characteristics of prediction error distributions to each subset of $\mathcal{C}$, we extend here the empirical approach described in Paragraph 5.4.1, by associating a collection of recent prediction errors to each of these subsets. As introduced in Equation (5.12), $\Omega_{t,k}$ is the set of all the past k-step ahead prediction errors up to time t. Define now $\Omega_{t,k}(\{(l, j_l)\})$ the subset of past prediction errors corresponding to the subset of forecast conditions $\mathcal{C}(\{(l, j_l)\})$:

$$\Omega_{t,k}(\{(l, j_l)\}) = \{\epsilon_{t-i+k/t-i} \in \Omega_{t,k} \mid c_{t-i,k} \in \mathcal{C}(\{(l, j_l)\})\}, \quad \forall j_l. \tag{5.24}$$

And finally, as we did in Equation (5.13), we can extract from each subset $\Omega_{t,k}(\{(l, j_l)\})$

a sample $S_{t,k}(\{(l,j_l)\})$ of size n containing the last n forecasting errors, but in *similar forecast conditions*:

$$S_{t,k}(\{(l,j_l)\}) = \{\epsilon_i \in \Omega_{t,k}(\{(l,j_l)\}),\ i=1,\ldots,n\}, \quad \forall j_l. \tag{5.25}$$

Therefore, each of the subsets $\mathcal{C}(\{(l,j_l)\})$ is characterized by its own empirical distribution function $\hat{F}^{\epsilon}_{t,k}(\{(l,j_l)\})$, drawn from a different sample of past errors. Note that $\hat{F}^{\epsilon}_{t,k}(\{(l,j_l)\})$ is a conditional distribution function since it is an estimate of the distribution function of prediction errors given that $c_{t,k}$ is an item of $\mathcal{C}(\{(l,j_l)\})$. This empirical distribution function puts probability $1/n$ on each element of $S_{t,k}(\{(l,j_l)\})$:

$$\hat{F}^{\epsilon}_{t+k/t}(\{(l,j_l)\}) \rightarrow \{\epsilon_i,\ \epsilon_i \in S_{t,k}(\{(l,j_l)\})\}, \tag{5.26}$$

5.4.3 The fuzzy inference model for producing conditional distribution functions

The previously described classification is the basis for deriving an empirical and distribution-free method that provides conditional prediction intervals, given particular forecast conditions. The choice of the influential variables, as well as the splitting of the sets of possible values into various subsets with different characteristics of related prediction error distributions, are the result of the expertise one has on the process of interest. It was explained in Paragraph 5.4.1 how to dress a point prediction $\hat{p}_{t+k/t}$ with an empirical distribution of prediction errors $\hat{F}^{\epsilon}_{t+k/t}$ for producing empirical distributions of wind generation $\hat{F}^{p}_{t+k/t}$. Hereafter, we develop a fuzzy inference model $h_f(c_{t,k})$ which gives conditional distributions of prediction errors $\hat{F}^{\epsilon,*}_{t+k/t}(c_{t,k})$ given the forecast condition $c_{t,k}$.

Fuzzy logic is an alternative paradigm to that of binary logic for which an event can only be associated to a true or false statement (and therefore 1 or 0). It considers instead that to each event can be associated a degree of truth, which is a continuous function between 0 and 1. For an introduction to the fuzzy logic theory, we refer to [91]. In the previous Paragraph, the set $\mathcal{C}$ of possible forecast conditions has been mapped with several subsets $\mathcal{C}(\{(l,j_l)\})$ related to different characteristics of the forecast uncertainty. Particularly, we have explained that a given subset $\mathcal{C}(\{(l,j_l)\})$ is defined as the association of the subsets $V_l^{j_l}$ $(l=1,\ldots,L)$ for the various considered input variables (Equation (5.22)). Here, we associate a fuzzy set $\mathcal{A}_l^{j_l}$ to each of these V-subsets. A fuzzy set is characterized by a membership function $m_l^{j_l}(v_{t,k}^l)$, which tells what the degree of truth of $v_{t,k}^l$ being an element of $V_l^{j_l}$ is:

$$m_l^{j_l} : v_{t,k}^l \rightarrow m_l^{j_l}(v_{t,k}^l) \in [0,1]. \tag{5.27}$$

The subset of forecast conditions $\mathcal{C}(\{(l,j_l)\})$ is defined as the association of the L subsets $V_l^{j_l}$. Therefore, the degree of truth of a given forecast condition $c_{t,k} = \{v_{t,k}^l\}_{l=1,\ldots,L}$ being an element of $\mathcal{C}(\{(l,j_l)\})$ is given by the product of the membership values for

every influential variable:

$$m(c_{t,k}, \{(l, j_l)\}) \;=\; m\left(c_{t,k} \in \mathcal{C}(\{(l, j_l)\})\right) \;=\; \prod_{l=1}^{L} m_l^{j_l}(v_{t,k}^l). \tag{5.28}$$

The basic element of the fuzzy inference model we develop here consists in fuzzy rules. Such a fuzzy rule can be expressed as

$$\text{“ } \textbf{IF} \quad v_{t,k}^1 \in \mathcal{D}(\mathcal{A}_1^{j_1}) \text{ and } \ldots \text{ and } v_{t,k}^L \in \mathcal{D}(\mathcal{A}_L^{j_L}) \quad \textbf{THEN} \quad \epsilon_{t+k/t} \sim F_{t,k}^{\epsilon}(\{(l, j_l)\}) \text{ ", } \tag{5.29}$$

where $\mathcal{D}(\mathcal{A}_l^{j_l})$ stands for the support of the fuzzy set $\mathcal{A}_l^{j_l}$. The 'IF' part is referred to as the premise of the rule, whereas the 'THEN' part is called the conclusion. Note that the above rule is equivalent to:

$$\text{“ } \textbf{IF} \quad c_{t,k} \in \mathcal{D}(\mathcal{A}_{\mathcal{C}}(\{(l, j_l)\})) \quad \textbf{THEN} \quad \epsilon_{t+k/t} \sim \hat{F}_{t,k}^{\epsilon}(\{(l, j_l)\}) \text{ ". } \tag{5.30}$$

where

$$\mathcal{D}(\mathcal{A}_{\mathcal{C}}(\{(l, j_l)\})) \;=\; \mathcal{D}(\mathcal{A}_1^{j_l}) \times \ldots \times \mathcal{D}(\mathcal{A}_L^{j_L}). \tag{5.31}$$

Actually, the rule (5.30) states that if the forecast condition $c_{t,k}$ can be considered as being an item of a given subset $\mathcal{C}(\{(l, j_l)\})$ of $\mathcal{C}$, then the prediction error $\epsilon_{t+k/t}$ follows the distribution $F_{t,k}^{\epsilon}(\{(l, j_l)\})$.

Then, a rule base is composed by rules similar that given by (5.30), which span all the possible subsets of $\mathcal{C}$. The number of fuzzy rules is hence given by the number of subsets N_s used to map the set of possible forecast conditions. For convenience, we associate an index i to each of the N_s subsets, and we introduce the function $\eta(i)$ that returns the $\{(l, j_l(i))\}_{l=1,\ldots,L}$ pairs that serve to identify the corresponding subset:

$$\eta \;:\; i \in \{1, \ldots, N_s\} \;\rightarrow\; (\{(l, j_l(i))\}_{l=1,\ldots,L}), \tag{5.32}$$

such that each of the $\{(l, j_l(i))\}_{l=1,\ldots,L}$ pairs is given by a unique value of i. Consequently, the i^{th} rule of the fuzzy rule base is of the form:

$$\text{“ } \textbf{IF} \quad c_{t,k} \in \mathcal{D}(\mathcal{A}_{\mathcal{C}}(\eta(i))) \quad \textbf{THEN} \quad \epsilon_{t+k/t} \sim F_{t,k}^{\epsilon}(\eta(i)) \text{ ", } \tag{5.33}$$

where

$$\mathcal{D}(\mathcal{A}_{\mathcal{C}}(\eta(i))) \;=\; \mathcal{D}(\mathcal{A}_1^{j_l(i)}) \times \ldots \times \mathcal{D}(\mathcal{A}_L^{j_L(i)}). \tag{5.34}$$

The inference procedure for the fuzzy logic model consists in applying the rule-base to the forecast condition $c_{t,k}$ in order to provide the overall conclusion as the weighted average of the conclusion of each rule. The weight w_i for each rule is given by the degree

of truth of the related premise, normalized by the sum of the weights for each rule:

$$w_i(c_{t,k}) = \frac{m(c_{t,k}, \eta(i))}{\sum_{i=1}^{N_s} m_{\eta(i)}(c_{t,k})}, \quad i = 1, \ldots, N_s, \tag{5.35}$$

with $m(c_{t,k}, \eta(i))$ defined by Equation (5.28).

By doing so, the fuzzy model tells what is the contribution of each of the $F^{\epsilon}_{t,k}(\eta(i))$ $(i = 1, \ldots, N_s)$ error distributions in the error distribution $F^{\epsilon}_{t,k}$ related to the current forecast condition $\hat{c}_{t,k}$. Finally, the fuzzy logic model can be written as

$$h_f : c_{t,k} \rightarrow \epsilon_{t+k/t} \sim F^{\epsilon,*}_{t,k} = \sum_{i=1}^{N_s} w_i(c_{t,k}).F^{\epsilon}_{t,k}(\eta(i)). \tag{5.36}$$

Let us draw an illustration of the fuzzy inference process by going back to the example of the above Paragraph, in which predicted power and forecast wind direction were considered as influential variables. For both input variables, the sets of possible values were split into two subsets. This would yield four samples of prediction errors $S_{t,k}((1,1),(2,1))$, $S_{t,k}((1,1),(2,2))$, $S_{t,k}((1,2),(2,1))$, $S_{t,k}((1,2),(2,2))$ related to different subsets of forecast conditions $\mathcal{C}((1,1),(2,1))$, $\mathcal{C}((1,1),(2,2))$, $\mathcal{C}((1,2),(2,1))$, $\mathcal{C}((1,2),(2,2))$, for a given time t and look-ahead time k. Moreover, the fuzzy logic model in that case would have a rule-base composed by four rules, one for each of the possible $\mathcal{C}$-subsets. Then, imagine that for given time t and horizon k the degrees of truth of the current forecast condition $c_{t,k}$ being part of the $\mathcal{C}$-subsets are evaluated to be respectively equal to 0.3, 0.5, 0.15 and 0.05. The fuzzy rule-base (5.36) then defines the corresponding distribution of prediction errors as:

$$\begin{aligned} \epsilon_{t+k/t} \sim F^{\epsilon,*}_{t,k} \quad = \quad & 0.3\ F^{\epsilon}_{t,k}((1,1),(2,1)) \ + \ 0.5\ F^{\epsilon}_{t,k}((1,1),(2,2)) \\ & + 0.15\ F^{\epsilon}_{t,k}((1,2),(2,1)) \ + \ 0.05\ F^{\epsilon}_{t,k}((1,2),(2,2)). \end{aligned} \tag{5.37}$$

5.4.4 Methods for combining error distributions

In the above Paragraph, we have developed a fuzzy inference model h_f that provides conditional distribution functions of prediction errors. Given a specific forecast condition $c_{t,k}$, it returns the distribution $F^{\epsilon,*}_{t,k}$ of prediction errors $\epsilon_{t+k/t}$ as a combination of several distributions, corresponding to different subsets of the forecast conditions.

Combining probability distributions is not a trivial task. Perhaps the area which is the most concerned with the probability-combination problem is the area of probabilistic risk analysis and decision science. It is often demanded to a panel of experts to provide their judgment on a particular event in the form of probability distributions. A decision maker has then to assimilate the various experts' judgments, which may be converging or conflicting. Hence, the corresponding probability distributions can have significantly different shapes, and, in a general manner, they cannot be seen as Gaussian or even symmetric. The assimilation procedure followed by the decision maker consists in summarizing the various experts' opinion in a single *combined probability distribution*. In the last decades, several methods have been developed for that purpose, either of the mathemat-

ical or of the behavioral types. These methods are reviewed by Clemen and Winkler in [17]. In the following, we describe two alternative approaches for combining probability distributions: the linear opinion pool and adapted resampling.

The linear opinion pool

An appealing approach to the aggregation of probability distributions is the *linear opinion pool*, which consists in saying that a combined distribution is the weighted average of the individual probability distributions, the weights being non-negative and summed to one [88]. One notices then that this is exactly what is given by the fuzzy logic inference model described by Equation (5.36): the probability distribution of errors is given as the weighted average of probability distributions for various forecast condition subsets. Genest and McConway [30] discussed the interpretation of the weights to be assigned to individual probability distributions. While it appears obvious that they relate to the confidence one may have in such or such distribution to give more information on the true effect, it is not evident how they should be calculated. In our case, the weights are derived from the fuzzy-inference model, which, for a given forecast condition $\hat{c}_{t,k}$, tells to what extent we expect the error distribution for each subsets of $\mathcal{C}$ to represent the actual error distribution. Moreover, we assume that by considering non-overlapping subsets of forecast conditions the related error distributions $F^{\epsilon}_{t,k}(\eta(i))$ $(i = 1 \ldots, N_s)$ can provide independent and relevant information on the true distribution.

In Paragraph 5.4.1 it was explained that an error distribution function could be approximated by its related empirical distribution function that puts an equal probability to every item of a sample of past errors. Straightforwardly, we approximate here each $F^{\epsilon}_{t,k}(\eta(i))$ by the related $\hat{F}^{\epsilon}_{t,k}(\eta(i))$ $(i = 1 \ldots, N_s)$. Consequently, an estimate of the distribution $F^{\epsilon,*}_{t,k}$ follows from Equation (5.36):

$$\hat{F}^{\epsilon,*}_{t,k}(c_{t,k}) = \sum_{i=1}^{N_s} w_i(c_{t,k}).\hat{F}^{\epsilon}_{t,k}(\eta(i)). \tag{5.38}$$

By gathering all the error sample $S_{t,k}(\eta(i))$ $(i = 1 \ldots, N_s)$ available at time t and for horizon k, we define $S^*_{t,k}$ such that

$$S^*_{t,k} = S_{t,k}(\eta(1)) \cup \ldots \cup S_{t,k}(\eta(N_s)). \tag{5.39}$$

Therefore, given that the size of the error sample $S_{t,k}(\eta(i))$ $(i = 1 \ldots, N_s)$ is set to n, $S^*_{t,k}$ is composed by $n.N_s$ elements. An estimate of the distribution $F^{\epsilon,*}_{t,k}$ is given by the discrete distribution that puts a probability $w_j = w_i(c_{t,k})/n$ to every element ϵ_j of $S^*_{t,k}$ that is originally an element of $S_{t,k}(\eta(i))$:

$$\hat{F}^{\epsilon,*}_{t,k}(c_{t,k}) \rightarrow \{\epsilon_j \in S^*_{t,k},\ \mathrm{P}(\epsilon_j \mid \epsilon_j \in S^*_{t,k} \cap S_{t,k}(\eta(i))) = w_j = w_i(c_{t,k})/n\}. \tag{5.40}$$

As in the previous developments, the predictive distribution of wind generation is constructed by associating the estimate of the distribution of prediction errors to the point

forecast itself:

$$\hat{F}^{p,*}_{t,k}(c_{t,k}) \rightarrow \{\hat{p}_{t+k/t} + \epsilon_j,\ \mathrm{P}(\epsilon_j \mid \epsilon_j \in S^*_{t,k} \cap S_{t,k}(\eta(i))) = w_j = w_i(c_{t,k})/n\}. \quad (5.41)$$

Note that $\hat{F}^{p,*}_{t,k}$ is now a continuous function of forecast conditions.

The cumulative distribution function $\hat{G}^{p,*}_{t,k}$ related to $\hat{F}^{p,*}_{t,k}$ has a slightly different form than that of Equation (5.14), since the items of $S^*_{t,k}$ do not have the same probabilities:

$$\hat{G}^{p,*}_{t,k}(\epsilon) = \sum_{j=0}^{n.N_s} w_j.\mathbf{1}_{\epsilon_j<\epsilon}, \quad (5.42)$$

where $\mathbf{1}_{\epsilon_j<\epsilon}$ takes the value 1 if $\epsilon_j < \epsilon$ and 0 otherwise.

However, $\hat{G}^{p,*}_{t,k}$ can be used similarly for estimating the lower and upper bounds of the central prediction interval $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$ with nominal coverage rate $(1-\alpha)$ by picking the quantiles with proportion respectively $(\alpha/2)$ and $(1-\alpha/2)$ of the predictive distribution $\hat{F}^{p,*}_{t+k/t}$:

$$\hat{L}^{(\alpha)}_{t+k/t} = \hat{p}_{t+k/t} + \left(\hat{G}^{p,*}_{t,k}\right)^{-1}(\alpha/2), \quad (5.43)$$

$$\hat{U}^{(\alpha)}_{t+k/t} = \hat{p}_{t+k/t} + \left(\hat{G}^{p,*}_{t,k}\right)^{-1}(1-\alpha/2). \quad (5.44)$$

The adapted resampling method

The aim of methods like resampling (or bootstrapping, following the terminology of its inventor Efron [25]) is to have a better idea of a population distribution parameter (e.g. its mean or standard deviation) by going through a representative sample a high number of times. This manipulation of the representative sample can serve to associate a measure of accuracy to the estimate of this population parameter. Actually, bootstrapping has also been considered in the forecasting literature for estimating prediction intervals associated to point forecasts (see Clements and Taylor [19], Grigoletto [36], or Reeves [82] among others). Such a method has the advantage of being non-parametric, but it needs to have access to the analytic model. This is not conceivable here, since the approach we aim at developing assumes that the point prediction method is a kind of black-box, and thus that we do not have access to the underlying model. Resampling is used here as an alternative to the linear opinion pool approach for estimating quantiles of combined probability distributions.

Write $S = \{\epsilon_j\}_{j=1,\ldots,n}$ a random sample from a probability distribution F. The observations ϵ_j $(j = 1,\ldots,n)$ are assumed to be i.i.d. (independent and identically distributed) F. Following Efron's terminology, the *plug-in estimate* of a parameter $\theta = h(F)$ is defined to be $\hat{\theta} = h(\hat{F})$. This means that we estimate the true parameter of F by applying the same function to the empirical distribution function $\hat{F}$. This is what we have done in Equations (5.16) and (5.17) for estimating the lower and upper bounds of the prediction intervals. The elements of S are used for setting up an estimate $\hat{G}$ of the cumulative distribution function associated to F.

Denote by $X = \{x_j\}_{j=1,\dots,n}$ a random sample that is i.i.d. $U[0,1]$. The theory of probabilities tells us that the sample $G^{-1}(X) = \{G^{-1}(x_j)\}_{j=1,\dots,n}$ is i.i.d. F. Then, the idea of resampling states that since $\hat{G}$ is an estimate of the true cumulative distribution associated to F, one can use it for drawing alternative samples that would lead to other empirical distribution functions of the true distribution F. In practice, this alternative sample $S^{(b)}$ $(b = 1, \dots, B)$ is called a *bootstrap sample* and is obtained by picking *randomly* and *with replacement* n values out of the original sample S. $\hat{\theta}^{(b)}$ is a *bootstrap replication* of the θ statistic. Since all the bootstrap replications are potential estimates of the true parameter θ, one can consider them for calculating the bias or standard deviation associated to the original estimate $\hat{\theta}$, or even confidence intervals.

Here, we propose to apply the idea of resampling for estimating a given parameter θ of a combined probability distribution, by having a slightly different interpretation of the combination given by the fuzzy inference model (5.36) than that of the linear opinion pool. Remember that the fuzzy inference model gives a weight to each of the N_s distributions $F^{\epsilon}_{t,k}(\eta(i))$. The distributions $F^{\epsilon}_{t,k}(\eta(i))$ can be approximated by the empirical distributions $\hat{F}^{\epsilon}_{t,k}(\eta(i))$. The linear opinion pool approach states that these weights can be seen as probabilities and that one can construct a combined distribution by associating these probabilities to each sample. The difference we introduce here is that these weights w_i $(i = 1, \dots, N_s)$ are to be used for defining the share of each of the representative samples of errors $S_{t,k}(\eta(i))$ for defining a representative sample drawn from the combined distribution. We will use that interpretation for creating B bootstrap sample $S^{(b)}_{t,k}$ and compute a bootstrap replication $\hat{\theta}^{(b)}$ for each of them. Given n the size of the error samples, a bootstrap sample $S^{(b)}_{t,k}$ (also of size n) is constructed as following:

$$S^{(b)}_{t,k} = \{S^{(b)}_{t,k}(\eta(i))\}_{i=1,\dots,N_s}, \tag{5.45}$$

such that

$$S^{(b)}_{t,k}(\eta(i)) = \{\epsilon_j \mid \epsilon_j \in S_{t,k}(\eta(i))\}_{j=1,\dots,w_i.n}, \quad i = 1, \dots, N_s, \tag{5.46}$$

where the items of $S^{(b)}_{t,k}(\eta(i))$ are picked randomly and with replacement from $S_{t,k}(\eta(i))$.

The parameters of interest are the quantiles of the combined probability distribution $\hat{F}^{\epsilon}_{t,k}$. Therefore, write $\hat{G}^{\epsilon,(b)}_{t,k}$ the cumulative distribution function associated to the empirical distribution function $\hat{F}^{\epsilon,(b)}_{t,k}$ (following the definition of Equation (5.14)). The bootstrap replications of the lower and upper bounds of the interval forecast $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$ with nominal coverage rate $(1-\alpha)$ are given by:

$$\hat{L}^{(\alpha)(b)}_{t+k/t} = \hat{p}_{t+k/t} + \left(\hat{G}^{\epsilon,(b)}_{t,k}\right)^{-1}(\alpha/2), \tag{5.47}$$

$$\hat{U}^{(\alpha)(b)}_{t+k/t} = \hat{p}_{t+k/t} + \left(\hat{G}^{\epsilon,(b)}_{t,k}\right)^{-1}(1-\alpha/2). \tag{5.48}$$

Finally, we approximate the bootstrap expectation by taking the mean of all the boot-

strap replications, in order to obtain an estimate of the interval limits:

$$\hat{L}^{(\alpha)}_{t+k/t} = \frac{1}{B}\sum_{b=1}^{B}\hat{L}^{(\alpha)(b)}_{t+k/t}, \tag{5.49}$$

$$\hat{U}^{(\alpha)}_{t+k/t} = \frac{1}{B}\sum_{b=1}^{B}\hat{U}^{(\alpha)(b)}_{t+k/t}. \tag{5.50}$$

Note that by constituting these B bootstrap samples, we actually use all the information included in the individual samples by drawing alternatives scenarios. Also, while Efron and Tibshirani (see [26], pp. 124-126) explain that the bootstrap expectation serves for calculating the bias associated to the original estimate of a distribution parameter from a single sample, it has a completely different meaning here, since we apply that form of resampling for a multi-sample problem. In the remaining of the document, this approach is referred to as *adapted resampling* owing to the similarities with the original resampling approach.

5.5 Application to the wind power forecasting problem

In this Section, we detail how the previously introduced methods can be straightforwardly applied to the specific case of the wind power forecasting problem, by describing a particular configuration that accounts for both the nonlinearity related to the level of forecast power and the one related to the level of forecast wind speed (owing to the cut-off risk). Therefore, following the notations used in the previous Section, let us consider two influential variables ($L = 2$):

$$v^1_{t,k} = \hat{p}_{t+k/t}, \quad v^1_{t,k} \in V_1, \tag{5.51}$$

$$v^2_{t,k} = \hat{u}_{t+k/t}, \quad v^2_{t,k} \in V_2. \tag{5.52}$$

The forecast condition $c_{t,k}$ at time t, for lead time $t+k$, is then given by the pair consisting of the forecast wind speed and predicted power values

$$c_{t,k} = \{v^1_{t,k}, v^2_{t,k}\} = \{\hat{p}_{t+k/t}, \hat{u}_{t+k/t}\}, \quad c_{t,k} \in \mathcal{C} = V_1 \times V_2. \tag{5.53}$$

To account first for the power curve effects, the set V_1 of possible power values is divided into three subsets ($J_1 = 3$), corresponding to the power ranges 'low' (V^1_1), 'medium' (V^2_1) and 'high' (V^3_1):

$$V_1 = V^1_1 \cup V^2_1 \cup V^3_1. \tag{5.54}$$

In parallel, V_2 is divided into two subsets ($J_2 = 2$), corresponding to the range of forecast wind speed values for which a cut-off event is not expected (V^1_2), and to the range of values for which a cut-off is probable (V^1_2):

$$V_2 = V^2_2 \cup V^2_2. \tag{5.55}$$

This constitutes six different subsets of the forecast conditions ($N_s = 6$):

$$\mathcal{C}_1 = V_1^1 \times V_2^1, \tag{5.56}$$
$$\mathcal{C}_2 = V_1^2 \times V_2^1, \tag{5.57}$$
$$\mathcal{C}_3 = V_1^3 \times V_2^1, \tag{5.58}$$
$$\mathcal{C}_4 = V_1^1 \times V_2^2, \tag{5.59}$$
$$\mathcal{C}_5 = V_2^2 \times V_2^2, \tag{5.60}$$
$$\mathcal{C}_6 = V_3^3 \times V_2^2. \tag{5.61}$$

However, if considering a theoretical power curve such as the one depicted in Figure 5.2 it appears unlikely that a cut-off event occurs when predicted power values are in the 'low' or 'medium' ranges, and that the possibility of a cut-off event is only dictated by the forecast wind speed, the three subsets formed with V_2^2 are grouped to form only one:

$$\mathcal{C}_{4+} = \mathcal{C}_4 \cup \mathcal{C}_5 \cup \mathcal{C}_6. \tag{5.62}$$

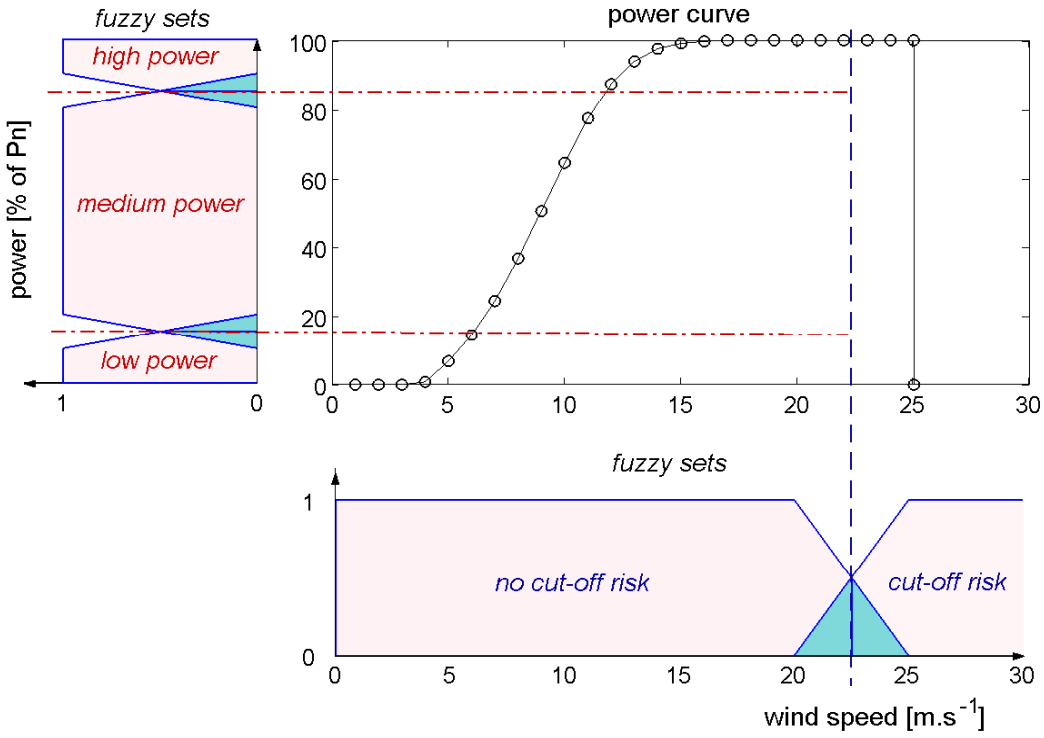

Figure 5.2: *Mapping of the forecast uncertainty introduced by the power curve. The range of possible predicted power values is divided into three ranges ('low', 'medium' and 'high'), to which are associated three trapezoidal fuzzy sets, in order to account for the nonlinearity introduced by the power variable. Similarly, the range of possible forecast wind speed values is divided into two ranges ('no cut-off risk' and 'cut-off risk'), owing to the nonlinearity introduced by the cut-off, to which are associated two trapezoidal fuzzy sets. This yields four zones of the power curve related to different characteristics of power prediction error distributions.*

Denote by $S^1_{t,k}$, $S^2_{t,k}$, $S^3_{t,k}$ and $S^{4+}_{t,k}$ the samples (of size n) of prediction errors corresponding to the four subsets of forecast conditions introduced above. At a given time t, each of these samples contain the last n k-step ahead prediction errors made by the point prediction approach in the forecast conditions defined respectively by $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ and $\mathcal{C}_{4+}$.

To every subsets of V_1 and V_2 are associated trapezoidal fuzzy sets. Figure 5.2 illustrates this mapping of a theoretical power curve into various zones corresponding to different characteristics of the prediction error distributions. Then, denote by $\mathcal{A}^i_{\mathcal{C}}$ the two-dimensional fuzzy sets related to the subset of forecast conditions $\mathcal{C}_i$, $i = 1, \ldots, 4+$. Each two-dimensional fuzzy set $\mathcal{A}^i_{\mathcal{C}}$ is characterized by its membership function $m(., i)$, $i = 1, \ldots, 4+$. The analytical form of these membership functions is not given here.

The fuzzy rule base inference model is composed by four fuzzy rules, which can be expressed as:

$$\text{“ IF } \quad c_{t,k} \in \mathcal{D}(\mathcal{A}^1_{\mathcal{C}}) \quad \text{THEN} \quad \epsilon_{t+k/t} \sim F^{\epsilon,1}_{t,k} \text{ ”}, \tag{5.63}$$

$$\text{“ IF } \quad c_{t,k} \in \mathcal{D}(\mathcal{A}^2_{\mathcal{C}}) \quad \text{THEN} \quad \epsilon_{t+k/t} \sim F^{\epsilon,2}_{t,k} \text{ ”}, \tag{5.64}$$

$$\text{“ IF } \quad c_{t,k} \in \mathcal{D}(\mathcal{A}^3_{\mathcal{C}}) \quad \text{THEN} \quad \epsilon_{t+k/t} \sim F^{\epsilon,3}_{t,k} \text{ ”}, \tag{5.65}$$

$$\text{“ IF } \quad c_{t,k} \in \mathcal{D}(\mathcal{A}^{4+}_{\mathcal{C}}) \quad \text{THEN} \quad \epsilon_{t+k/t} \sim F^{\epsilon,4+}_{t,k} \text{ ”}, \tag{5.66}$$

where $F^{\epsilon,i}_{t,k}$ is the empirical probability distribution that puts probability $1/n$ on each element of $S^i_{t,k}$, $i = 1, \ldots, 4+$. For instance, the first rule (given by (5.63)) states that if predicted power $\hat{p}_{t+k/t}$ is in the 'low' range and forecast wind speed $\hat{u}_{t+k/t}$ is in the 'no cut-off' range, prediction errors $e_{t+k/t}$ for that look-ahead time are distributed $F^{\epsilon,i}_{t,k}$.

The fuzzy inference model, which gives the conditional distributions of prediction errors as a function of the forecasts conditions, can thus be written as:

$$h_f \; : \; c_{t,k} \; \rightarrow \; \epsilon_{t+k/t} \sim F^{\epsilon,*}_{t,k} = \sum_{i=1}^{4+} w_i(c_{t,k}).F^{\epsilon,i}_{t,k}, \tag{5.67}$$

where the weights w_i of each of the fuzzy rules are calculated as following:

$$w_i(c_{t,k}) \; = \; \frac{m(c_{t,k}, i)}{\sum_{i=1}^{4+} m(c_{t,k}, i)}, \quad i = 1, \ldots, 4 + . \tag{5.68}$$

Let us now imagine an operational wind power forecasting application in which point predictions are produced from a state-of-the-art method (say one of the methods considered in the Anemos project). Denote by k_{max} the forecast length. The size n of the error samples is defined by the end-user, as well as the nominal coverage rate $(1 - \alpha)$ of the interval forecasts. The Algorithm 5.1 describes the steps for the estimation of prediction intervals of wind generation at prediction time t. In a first stage, one retrieves the power measure p_t at time t and the series of predictions $\hat{p}_{t+k/t}$ $k = 1, \ldots, k_{max}$ produced from the point forecasting method. The power measure is used for calculating the errors $e_{t/t-k}$ related to the predictions $\hat{p}_{t/t-k}$ issued in the past for time t. It is thus necessary to store the series of predictions for a time period equal to the forecast length of the considered point prediction method. This is also valid for the case of the considered influential variables.

Forecast conditions $\hat{c}_{t-k,k}$ related to $e_{t/t-k}$ are determined in order to decide to which error samples the prediction errors $\hat{e}_{t/t-k}$ belong to. These samples are then updated by discarding the oldest error value and by adding the new one as it becomes available[5]. This makes the method *adaptive*, since it always considers the most recent information on the process. It can accommodate temporal modifications of the characteristics of prediction error distributions, owing to the season of the year, changes in the wind farm environment, etc. The fuzzy inference model (Equation (5.67)) is used independently for each prediction horizon, for determining how to estimate conditional prediction error distributions $F^{\epsilon,*}_{t,k}$ given the forecast conditions $c_{t,k}$. The prediction intervals $\hat{\mathrm{I}}^{(\alpha)}_{t+k/k}$ with nominal coverage rate $(1-\alpha)$ are finally estimated by applying either the linear opinion pool or the resampling approach for the combination of probability distributions given by the fuzzy inference model.

Algorithm 5.1: *The necessary steps at time t for producing the empirical and distribution-free prediction intervals of wind generation*

step 1.	Retrieve the power measure p_t for time t
step 2.	Retrieve and store the power predictions $\hat{p}_{t+k/t}$, $k > 0$, provided by a point prediction method, as well as related influential variables values
step 3.	Calculate the prediction errors $e_{t/t-k}$ related the power predictions $\hat{p}_{t/t-k}$, $k > 0$, issued at time $t-k$ for time t
step 4.	Update the relevant error samples given the forecast condition $c_{t-k,k}$ related to the prediction error $e_{t/t-k}$
step 5.	For each look-ahead time k, use the fuzzy inference model given by Equation (5.67) to determine the distribution of prediction errors $F^{\epsilon}_{t,k}$ given the forecast conditions $c_{t,k}$
step 6.	For each look-ahead time k, apply either the linear opinion pool or the adapted resampling approach for estimating the bounds of the prediction intervals of wind generation $\hat{\mathrm{I}}^{(\alpha)}_{t+k/t}$

The proposed methods for the estimation of prediction intervals of wind generation has originally been developed for online application. In Appendix B, we detail the characteristics of the module we have implemented and which is integrated in the ANEMOS prediction platform.

5.6 Discussion on operational aspects

What type of prediction intervals?

The above methods permit to estimate predictive distributions of wind generation. We

[5] At the beginning of the application, error samples are empty. But, as new predictions are provided and related power measures made available, these samples are filled and updated. Even if the number of items in each sample has not reached n, it is possible to apply the previously described methods by modifying the necessary steps, i.e. by considering the number of available errors instead of the required n elements. After a certain time of operation (a minimum of $n.N_s$ forecasting steps), all the samples attain their defined size.

proposed to summarize the uncertainty information by quoting prediction intervals, which consist of two particular quantiles of these distributions. Actually, even by assigning a certain nominal coverage rate, the resulting intervals with that pre-assigned probability can be intervals around the mean, the median, or intervals with the shortest length for instance. It was already explained in Section 5.2 that it was more appropriate to provide central prediction intervals than intervals around the distribution mean, since they correspond to quantiles with known proportions. An alternative described by Hyndman [44] is to provide highest-density interval forecasts, which are defined as intervals with the shortest length given that any point inside the intervals has a probability density at least as large as every point outside the intervals. They can be of practical interest when working with non-normal and multimodal distributions. But, again, the endpoints of these intervals do not correspond to quantiles with known proportions. For instance, if focusing on the lower bound of intervals with a 80% degree of confidence, it may correspond to a quantile with proportion 0.05 for a given forecast and then to a quantile with proportion 0.15 for the following one. Thus, the risk of the true effect lying below that lower bound will change for every prediction. This does not appear appropriate from an operational point of view. Therefore, we will focus on central prediction intervals only.

Choice of an optimal coverage rate

An important question concerning the intervals arises: how to choose an optimal nominal coverage rate? When this pre-assigned probability is higher than 90%, intervals can be 'embarrassingly' wide, because they will contain extreme prediction errors (or even outliers). Working with high-coverage intervals means that we are aiming at modeling the very tails of the error distributions. Thus, the robustness of the uncertainty estimation methods becomes a critical aspect. However, if one defines lower pre-assigned probabilities (50% for instance), intervals will be much more narrow and more robust with respect to extreme prediction errors. But, this would translate to future observations being equally likely to lie inside or outside these bounds. In both cases, prediction intervals appear hard to handle and that is why an intermediate degree of confidence (75-85%) seems to be a good compromise [15].

Marginal or simultaneous prediction intervals

Moreover, the fact that prediction intervals are designed for multi-step ahead forecasts imposes to define what is the real required degree of confidence. As a matter of fact, there is a difference between a nominal coverage rate defined for each predicted value and a nominal coverage rate that would be defined over the whole forecast length. For instance, if a 85% degree of confidence is required for one-day ahead hourly predictions, the former corresponds to "each of the 24 intervals will contain the true value 85% of the times" (referred to as *marginal intervals* in the forecasting literature, though the term pointwise may be more appropriate), while the latter translates to "the 24 intervals will contain all the 24 true values 85% of the times" (referred to as *simultaneous intervals* in the literature). The second way of reasoning is obviously much more restrictive and seems less applicable in our case. As we explained in previous Sections, the method for interval estimation is applied separately for every look-ahead time. Therefore, the

observed confidence will be verified accordingly. For a more thorough discussion about multi-step ahead prediction intervals, we refer to Chan et al. [14] and Ravishankar et al. [81].

Multiple intervals for providing predictive distributions of wind generation

Instead of focusing on a particular nominal coverage rate, it seems that producing a number of prediction intervals for a range of nominal coverage rates would be a better solution. This would allow one to build the whole probability distribution of expected wind generation for each look-ahead time. As explained in Section 5.4, this may involve the development of several models e.g. if considering quantile regression methods, but with the methodology described above, it is not more computationally expensive to estimate one or thirty quantiles. Wind power forecast users may request not only a single interval forecast but also predictive distributions of future wind generation. Indeed, the decision-making methods appearing in the literature need a complete approximation of the density function for providing an optimal management [23] or trading strategy [5, 28]. Therefore, we will consider that when possible, several interval forecasts with various degrees of confidence should be provided.

5.7 A non-parametric framework for the evaluation of prediction intervals

Evaluating probabilistic forecasts (either density or interval forecasts) is more complicated than evaluating deterministic ones. When it is easy to say that a point forecast is false because the deviation between the predicted and the real values is of practical magnitude, an individual probabilistic forecast cannot be deemed as incorrect [65]. Indeed, when an interval forecast states there is a 90% confidence that expected power generation (for a given horizon) will be between 100 and 250kW and that the actual outcome equals 90kW, how to tell if this case should be part or not of the 10% of cases for which intervals miss?

In this Section, our aim is to describe what the required properties of interval forecasts are, and how they can be evaluated in terms of their statistical performance. For that purpose, we present relevant skill scores and diagrams that were introduced in the statistical and meteorological literature. We consider here a non-parametric framework that is suitable for evaluating either intervals or series of quantiles. Moreover, in the following, all criteria are evaluated as a function of the look-ahead time, or as an average over the forecast length. If the evaluation set is large enough, it would also be appropriate to assess the skill of probabilistic forecasting methods as a function of some other parameters (e.g. level of power).

5.7.1 Required properties for interval forecasts

Prediction intervals are associated to a probability, which is their nominal coverage rate. The first requirement for interval forecasts is that their empirical coverage should be close

to the nominal one. Actually, if considering infinite series of interval forecasts, that empirical coverage should exactly equal the pre-assigned probability. That first property is referred to as *reliability* or *calibration* in the literature [2, 18, 65].

Besides this first requirement, it is necessary that prediction intervals provide a situation-dependent assessment of the forecast uncertainty. Their size should then vary depending on various external conditions. For the example of wind prediction, it is intuitively expected that prediction intervals (for a given nominal coverage rate) should not have the same size when predicted wind speed equals zero and when it is near cut-off speed. The most simple type of intervals is constant-size intervals (e.g. produced from climatology). Advanced methods for their estimation are expected to produce variable-size intervals. This property is commonly named *sharpness* or *resolution* of the intervals [2, 65]. Note that here, we will introduce a nuance between sharpness and resolution: the former will relate to the average size of intervals while the latter will be associated to their size variability.

Actually, the traditional view of interval forecast evaluation, which mainly comes from the econometric forecasting community, is based on the testing of correct conditional coverage. This means intervals have to be unconditionally reliable, and independent (see for instance [16], or [18] ch. 3)). However, in the case of wind power forecasting, we know there exists a correlation among forecasting errors (at least for short time-lags) [70]. Thus, we do not expect prediction intervals to be independent. Then, it appears preferable to develop an evaluation framework that is based on an alternative paradigm. We propose to consider reliability as a primary requirement and then sharpness and resolution as an added value. It should be noted here that reliability can be increased by using some re-calibration methods (e.g. conditional parametric models [74] or smoothed bootstrap [38]), while sharpness/resolution cannot be enhanced with post-processing procedures. This second aspect is the inherent (and invariant) ability of a probabilistic forecasting method to distinctly resolve future events [90].

5.7.2 Methods for the evaluation of prediction intervals

The following methods focus on the evaluation of predictive quantiles or prediction intervals of wind generation in a hierarchical manner: reliability has to be assessed first, followed by a study of sharpness and then resolution. A skill score is introduced in a second stage, which allows one to directly assess the overall quality of these predictions.

The indicator variable

Before going further with the evaluation of interval forecasts, it is necessary to introduce the *indicator variable* $\mathcal{I}_{t,k}^{(\alpha)}$ (following the definition by Christoffersen [16]), which is defined for a prediction made at time t and for the horizon k as follows

$$\mathcal{I}_{t,k}^{(\alpha)} = \mathbf{1}_{p_{t+k} \in \hat{\mathrm{I}}_{t+k/t}^{(\alpha)}} = \begin{cases} 1, & \text{if} \quad p_{t+k} \in [\hat{L}_{t+k/t}^{(\alpha)}, \hat{U}_{t+k/t}^{(\alpha)}] \\ 0, & \text{otherwise} \end{cases} . \tag{5.69}$$

This indicator variable tells if the actual outcome p_{t+k} at time $t+k$ lies ("*hit*") or not ("*miss*") in the prediction interval estimated for that lead time. We would like to mention

that this definition of the indicator variable can be easily adapted when working with quantiles of a probabilistic distribution. Indeed, the value of p_{t+k} lying or not inside the interval is replaced by the test of p_{t+k} being below or above the estimated quantile $\hat{r}^{(\alpha)}_{t+k/t}$. Then, $\mathcal{I}^{(\alpha)}_{t,k}$ can alternatively be defined with

$$\mathcal{I}^{(\alpha)}_{t,k} = \mathbf{1}_{p_{t+k} \leq \hat{r}^{(\alpha)}_{t+k/t}} = \begin{cases} 1, & \text{if} \quad p_{t+k} \leq \hat{r}^{(\alpha)}_{t+k/t} \\ 0, & \text{otherwise} \end{cases} . \tag{5.70}$$

Let then define as $n^{(\alpha)}_{k,1}$ the sum of hits and $n^{(\alpha)}_{k,0}$ the sum of misses (for a given horizon k) over the N_T realizations:

$$n^{(\alpha)}_{k,1} = \#\{\mathcal{I}^{(\alpha)}_{t,k} = 1\} = \sum_{t=1}^{N_T} \mathcal{I}^{(\alpha)}_{t,k}, \tag{5.71}$$

$$n^{(\alpha)}_{k,0} = \#\{\mathcal{I}^{(\alpha)}_{t,k} = 0\} = N_T - n^{(\alpha)}_{k,1}. \tag{5.72}$$

It is by studying the series of indicator variable $\{\mathcal{I}^{(\alpha)}_{t,k}, t = 1, \ldots, N_T\}$ over the test set that we will assess the reliability and overall skill of interval forecasts.

Reliability

The easiest way to check the reliability of interval forecasts is to compare their empirical coverage to the nominal one (i.e. the required probability (1-α)). An estimation $\hat{a}^{(\alpha)}_k$ of the actual coverage $a^{(\alpha)}_k$, for a given horizon k, is obtained by calculating the mean of the $\{\mathcal{I}^{(\alpha)}_{t,k}\}_{t=1,\ldots,N_T}$ time-series over the test set:

$$\hat{a}^{(\alpha)}_k = \frac{1}{N_T} \sum_{t=1}^{N_T} \mathcal{I}^{(\alpha)}_{t,k} = \frac{n^{(\alpha)}_{k,1}}{n^{(\alpha)}_{k,0} + n^{(\alpha)}_{k,1}}. \tag{5.73}$$

This standard measure for evaluating prediction intervals' reliability was first proposed by Ballie et al. [3] and by McNees [66]. This is the idea used in *reliability diagrams* which give the empirical coverage versus the nominal coverage for various nominal coverage rates. The closer to the diagonal the better. They can alternatively be depicted as the deviation from the 'perfect reliability' case for which empirical coverage would equal the nominal one (calculated as the difference between these two quantities). This idea is similar to the use of Probability Integral Transform histograms as proposed by Gneiting et al. [33], except that reliability diagrams directly provide that additional information about the magnitude of the deviations from the 'perfect reliability' case.

Reliability diagrams allow one to summarize the calibration assessment of several quantiles or intervals and thus to see at one glance if a given method tends to systematically underestimate (or overestimate) the uncertainty. Figure 5.3 shows an example of a reliability diagram for the evaluation of a given estimation method of wind power predictive distributions. Deviations from the 'perfect reliability' case are given as a function of the quantile nominal proportions, as an average over the forecast length. Here, one

notices a rather good calibration of the method since deviations are lower than 2%. However, the fact that quantiles are slightly overestimated for proportions lower than 0.5 and slightly underestimated for proportions above that value indicates that corresponding predictive distributions are a bit too narrow.

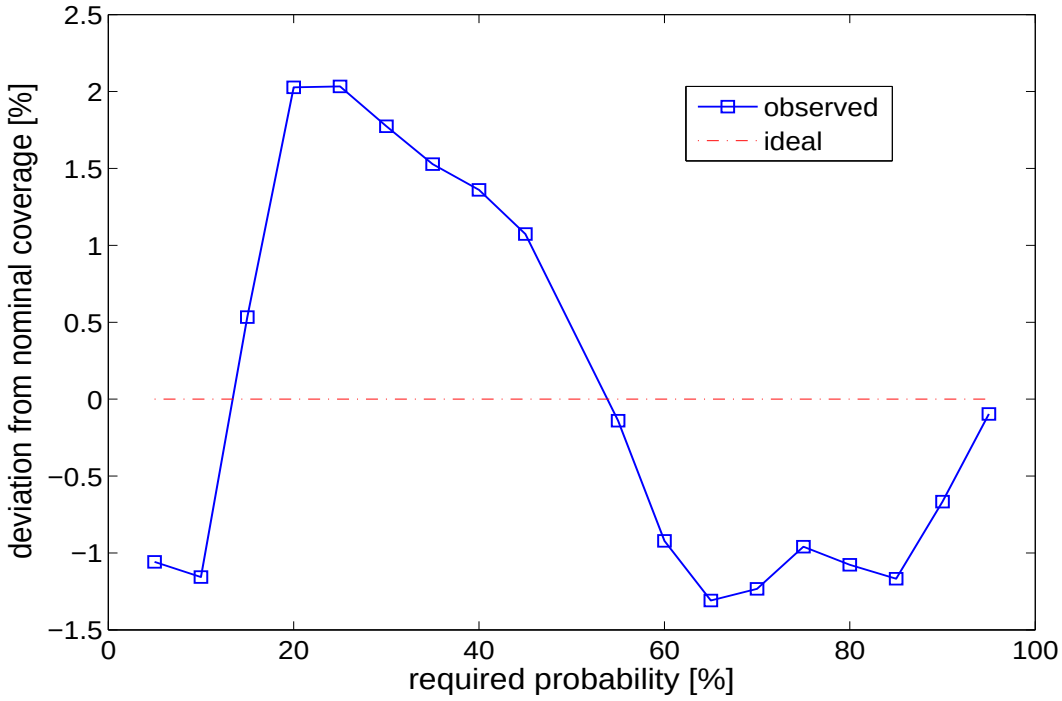

Figure 5.3: *Example of a reliability diagram depicting deviations as a function of the nominal coverage rate, for the reliability evaluation of a method providing probabilistic forecasts of wind generation.*

Using that kind of comparison between the nominal and empirical coverage introduces subjectivity in the evaluation: the decision of whether the intervals have correct coverage or not is left to the analyst. This is why a more objective framework based on hypothesis testing has been introduced in the forecasting literature (mainly in econometric forecasting). For instance, Christoffersen [16] proposed a likelihood ratio χ^2-test for evaluating the unconditional coverage of interval forecasts of economic variables, accompanied by another test of independence. In the area of wind generation forecasting, Bremnes [10] recently used a Pearson χ^2-test for evaluating the reliability of the quantiles produced from a local quantile regression approach. However, χ^2-tests rely on an independence assumption regarding the sample data. Owing to the correlation of wind power forecasting errors, it is expected that series of interval hits and misses can come clustered together in a time-dependent fashion. This actually means that independence of the indicator variable sequence cannot be assumed in our case (except if independence is proven in a prior analysis). In such cases, serial correlation invalidates the significance level of hypothesis tests. In general, it is known that statistical hypothesis tests cannot be directly applied for assessing the reliability of probabilistic forecasts due to the either

serial or spatial correlation structures [39].

Sharpness and Resolution

When dealing with sharpness or resolution, focus is given to the size of prediction intervals, or in a more general manner to the shape of predictive distributions. In the meteorological literature, the sharpness of probabilistic forecasts correspond to the ability of these forecasts to deviate from the climatological mean probabilities, whereas resolution stands for the ability of providing different conditional probability distributions $q(p|\hat{p})$ depending on the level of the predictand. For probabilistic forecasts with perfect reliability, these two notions are equivalent [90]. Here, we introduce a slightly different view of these two aspects. Given that the reliability of probabilistic forecasts is assessed in a prior analysis, we then propose to study the evolution of the shape of probabilistic distributions. Distributions that are narrower should be rewarded, since it will increase their value in a decision-making context. This is what we will regard as the sharpness of probabilistic forecasts. And, if rival probabilistic prediction methods produce distributions with a similar sharpness, then distributions whose shape exhibits larger variations over the evaluation period, hence showing a better ability for discriminating among future events, should be preferred. This is in line with the resolution aspect defined in the meteorological literature.

Define

$$\delta_{t,k}^{(\alpha)} = \hat{U}_{t+k/t}^{(\alpha)} - \hat{L}_{t+k/t}^{(\alpha)} = \hat{r}_{t+k/t}^{(1-\alpha/2)} - \hat{r}_{t+k/t}^{(\alpha/2)} \tag{5.74}$$

the size of the central interval forecast (with pre-assigned probability $(1-\alpha)$) estimated at time t for lead time $t+k$.

If two uncertainty estimation methods provide intervals at an acceptable level of reliability, we explained that it is the method that yields the narrowest intervals that is to be preferred. Here, the sharpness aspect is evaluated by calculating the average size $\bar{\delta}_k^{(\alpha)}$ of the prediction intervals for a given horizon k:

$$\bar{\delta}_k^{(\alpha)} = \frac{1}{N_T}\sum_{t=1}^{N_T}\delta_{t,k}^{(\alpha)} = \frac{1}{N_T}\sum_{t=1}^{N_T}\left(\hat{r}_{t+k/t}^{(1-\alpha/2)} - \hat{r}_{t+k/t}^{(\alpha/2)}\right). \tag{5.75}$$

Both Bremnes [10] and Nielsen et al. [73] used such a measure for evaluating the sharpness of the their probabilistic forecasts as a function of the horizon. When focusing on the distance between the quantiles for proportions 0.25 and 0.75 (i.e. the quartiles), this measure is commonly known as the inter-quartile range. However, since in a non-parametric framework probabilistic forecasts may consist in a set of prediction intervals, it would be interesting not to focus only on these two particular quantiles but also to look at the size of intervals corresponding to the very central and to the tail parts of the predictive distributions — say $\bar{\delta}_k^{(0.8)}$ and $\bar{\delta}_k^{(0.2)}$ for instance, which are the average size of the respectively 20%- and 80%-confidence central intervals.

In parallel, the resolution concept stands for the ability of providing a situation-

dependent assessment of the uncertainty. If two approaches have similar sharpness, then a higher resolution translates to a higher quality of related interval forecasts. It is not possible to directly verify that property, though we can study the variation in size of the intervals by using an appropriate summary statistic such as the standard deviation $\sigma_k^{(\alpha)}$ of the interval size (for a given horizon k and nominal coverage rate $(1-\alpha)$), where

$$\sigma_k^{(\alpha)} = \left[\frac{1}{N_T - 1} \sum_{t=1}^{N_T} \left(\delta_{t,k}^{(\alpha)} - \bar{\delta}_k^{(\alpha)} \right)^2 \right]^{\frac{1}{2}}. \tag{5.76}$$

Because of the nonlinear and conditionally heteroskedastic nature of the wind generation process, the forecast uncertainty is highly variable and it is thus expected that the interval size also greatly varies.

Finally, δ-diagrams and σ-diagrams, which give respectively $\bar{\delta}_k^{(\alpha)}$ and $\sigma_k^{(\alpha)}$ as a function of the nominal coverage rate for a given look-ahead time k (or over the forecast length), permit to better visualize the shape (and shape variations) of predictive distributions. We will underline the interest of such diagnostic tools in the following Section.

Defining a unique skill score

As for point-forecast verification, it is often demanded that a unique skill score would give the whole information on a given method performance. Such a measure would be given by scoring rules that associate a single numerical value $\mathrm{Sc}(\hat{q}, p)$ to a predictive distribution $\hat{q}$ if the event p materializes. Then, we can define as

$$\mathrm{Sc}(\hat{q}', \hat{q}) = \int \mathrm{Sc}(\hat{q}', p)\hat{q}(p)dp \tag{5.77}$$

the score under $\hat{q}$ when the predictive distribution is $\hat{q}'$.

A scoring rule should reward a forecaster that expresses his true beliefs. It is said to be *proper* if it does so. One remembers here that Murphy [68] referred to that aspect as the forecast *consistency* and stated that a forecast (probabilistic or not) should correspond to the forecaster's judgment. If we assume that a forecaster wishes to maximize his skill score over an evaluation set, then a scoring rule is said to be proper if for any two predictive distributions $\hat{q}$ and $\hat{q}'$ we have

$$\mathrm{Sc}(\hat{q}', \hat{q}) \leq \mathrm{Sc}(\hat{q}, \hat{q}), \quad \forall \hat{q}, \hat{q}'. \tag{5.78}$$

The scoring rule Sc is said to be strictly proper if Equation (5.78) holds with equality if and only if $\hat{q}' = \hat{q}$. Hence, if $\hat{q}$ corresponds to the forecaster's judgment, it is by quoting this particular predictive distribution that he will maximize his skill score.

If we consider that a predictive distribution $\hat{q}$ is characterized by its quantiles $\hat{\mathbf{r}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_l\}$ at levels $\alpha_1, \alpha_2, \ldots, \alpha_l$, Gneiting et Raftery [34] recently showed that any

scoring rule of the form

$$\mathrm{Sc}(\hat{\mathbf{r}}, p) = \sum_{i=1}^{l} \big(\alpha_i s_i(r_i) + (s_i(p) - s_i(r_i))\mathcal{I}^{(\alpha_i)} + f(p)\big), \tag{5.79}$$

with $\mathcal{I}^{(\alpha_i)}$ the indicator variable (for the quantile with proportion α_i) introduced above, s_i non-decreasing functions and f arbitrary, was proper for evaluating this set of quantiles. Here $\mathrm{Sc}(\hat{\mathbf{r}}, p)$ is a positively rewarding score: a higher score value stands for an higher skill. The specific case of central prediction intervals corresponds to the case where only two quantiles are quoted (cf. Equations (5.4) and (5.5)). Note that for a unique quantile, the scoring rule given by Equation (5.79) generalizes the loss functions considered in quantile regression [73] and local quantile regression [10].

Actually, Gneiting and Raftery [34] also noticed that for the specific case of central prediction intervals with nominal coverage rate $(1-\alpha)$, by putting $\alpha_1 = \alpha/2$ and $\alpha_2 = 1-\alpha/2$, $s_i(p) = 4p$, $(i = 1, 2)$, and $f(p) = -2p$, one retrieves an interval score that has already been proposed by Winkler [93]. Such an interval score $\mathrm{Sc}_{t,k}^{(\alpha)}$ used for evaluating the interval $\hat{\mathrm{I}}_{t+k/t}^{(\alpha)}$ has the following form:

$$\mathrm{Sc}_{t,k}^{(\alpha)} = \begin{cases} -2\alpha\delta_{t,k}^{(\alpha)} - 4(\hat{L}_{t+k/t}^{(\alpha)} - p_{t+k}), & \text{if} \quad p_{t+k} < \hat{L}_{t+k/t}^{(\alpha)} \\ -2\alpha\delta_{t,k}^{(\alpha)}, & \text{if} \quad p_{t+k} \in \hat{\mathrm{I}}_{t+k/t}^{(\alpha)} \\ -2\alpha\delta_{t,k}^{(\alpha)} - 4(p_{t+k} - \hat{U}_{t+k/t}^{(\alpha)}), & \text{if} \quad p_{t+k} > \hat{U}_{t+k/t}^{(\alpha)} \end{cases}, \tag{5.80}$$

where $\delta_{t,k}^{(\alpha)}$ is the size of the interval forecast $\hat{\mathrm{I}}_{t+k/t}^{(\alpha)}$ as defined in Equation (5.74).

This score is appealing since it considers the size of the intervals (by rewarding tight intervals) and gives a penalty if the observation does not lie inside the estimated interval. The score is calculated at each prediction time and then averaged over the test set in order to obtain the final score value $\mathrm{Sc}_{k}^{(\alpha)}$ for every horizon k

$$\mathrm{Sc}_{k}^{(\alpha)} = \frac{1}{N_T} \sum_{t=1}^{N_T} \mathrm{Sc}_{t,k}^{(\alpha)}. \tag{5.81}$$

Using a unique proper skill score allows one to compare the overall skill of rival approaches, since scoring rules such as the one given by Equation (5.79) encompass all the aspects of probabilistic forecast evaluation. It can also be utilized as a criterion for optimizing the parameters of a given quantile estimation method. However, a unique score does not tell what are the contributions of reliability or sharpness/resolution to the skill (or to the lack of skill)[6]. Though, if reliability is verified in a prior analysis, relying on a skill score permits to carry out an assessment of all the remaining aspects, namely sharpness and resolution.

[6]This has already been stated by Roulston et al. [83] when introducing the 'ignorance score', which despites its many justifications and properties has no ability to tell why a given method is better than another.

5.8 Results

The evaluation framework introduced in Section 5.7.2 is applied here for assessing the quality of the prediction intervals produced from both the linear opinion pool and the adapted resampling method. For that purpose, we have selected three statistical prediction methods considered in the Anemos project, that are denoted by M1, M2 and M3, and their application to the Tunø Knob and Klim case-studies. By selecting these methods that provide predictions every hour, there are more available data for evaluating interval forecasts. The Klim and Tunø Knob test cases consists of respectively 18943 and 3220 series of wind power point predictions and associated interval forecasts. In this Section, we assess the skill of the proposed interval estimation methods by showing and commenting on some selected results from the full verification procedure. Note that we do not consider any benchmark intervals based on an assumption about the shape of error distributions. We have already demonstrated the superiority of the proposed distribution-free approach against Box-Jenkins intervals [78] and also against intervals derived from the assumption that predictive distributions of wind generation can be modeled with β-distribution [79].

Regarding the mapping of the forecast uncertainty, since the available dataset only prove very few occurrences of cut-off events, it has not appeared appropriate to consider the nonlinearity introduced by the cut-off risk. The effects of the level of predicted power on the uncertainty are significant. Though, we do no know what is the influence of meteorological variables e.g. wind direction on prediction error distributions. We assume here that these effects can be neglected since there is no evidence in the literature of their impact on forecast uncertainty. Therefore, the mapping of the forecast conditions only concerns the range of possible predicted power values.

Figure 5.4 depicts an episode consisting in a set of wind power predictions provided by M2, issued on the 28th March 2003 at 10:00, for the Tunø Knob wind farm. The related power measures are also shown. Moreover, a set of interval forecasts produced from the adapted resampling method is associated to the point predictions, in the form of a fan chart. The nominal coverage rates for these intervals were set to 10, 20, ..., 90%. This illustrative example does not have any statistical value for assessing the quality of the intervals, but serves instead to show some of the nice properties of the designed approach.

When describing the uncertainty estimation methods, we explained that these methods were non-parametric (i.e. the intervals are estimated without assuming a specified distribution), and that this would permit to produce asymmetric prediction intervals. From the example of Figure 5.4, one clearly sees that interval forecasts are not symmetric around the point predictions. Also, one verifies a comment we made in Section 5.2: since intervals are central prediction intervals, they are centered around the median of the predictive distributions of wind generation and hence do not necessarily cover the point predictions themselves (which in turn are estimates of the mean of these distributions). Therefore, when the asymmetry of error distribution is more pronounced, for low and high predicted power values for instance, the difference between the center of the intervals and the point prediction is higher. This is clear here for horizons between 35-

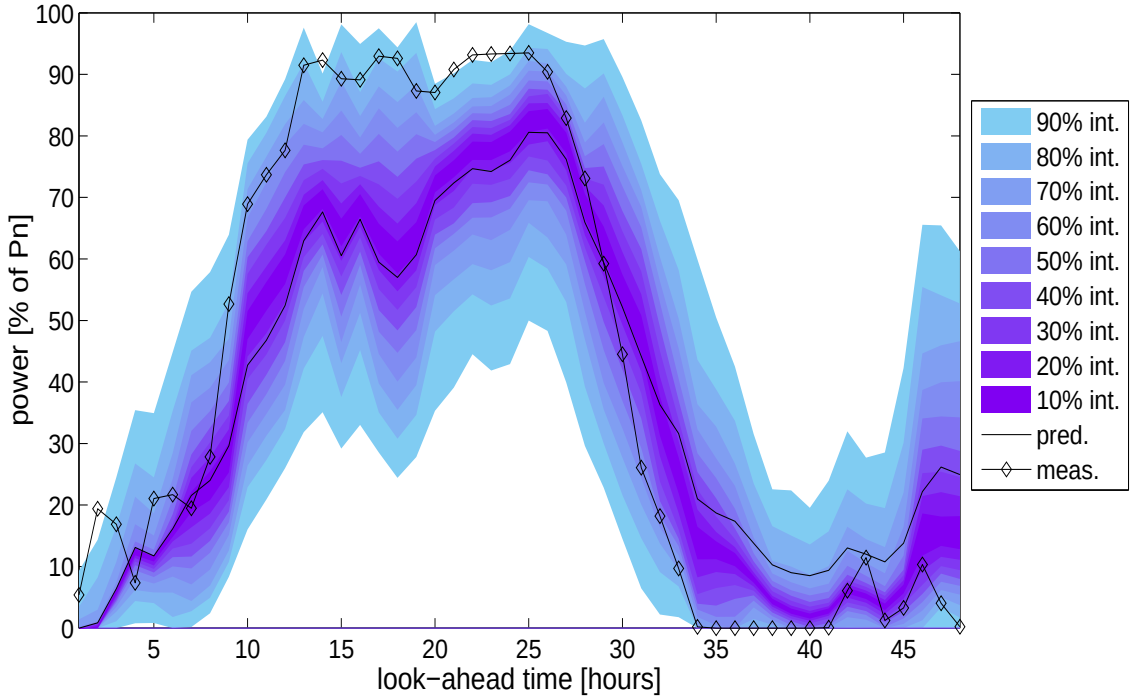

Figure 5.4: *Example of wind power point predictions associated with a set of interval forecasts. The point predictions are given by M2 and the central interval forecasts are estimated consequently with the adapted resampling method. Nominal coverage rates range from 10 to 90%. These sets of predictions and intervals were issued on the 28th March 2003 at 10:00, for the Tunø Knob wind farm.*

and 45-hour ahead. Note that the developed methods for estimating interval forecasts can be straightforwardly adapted if one wants to build prediction-centered intervals — this is done by considering separately distributions of positive and negative prediction errors.

Moreover, the effects of both the lead time and the level of predicted power can be seen from the Figure. Prediction intervals are very tight for the very first horizons, owing to the low level of predicted power and also because it is easier to predict for short-range horizons with statistical methods. Then, they get rather large when predicted power is in the medium-range: the forecast uncertainty is higher in such a case. Finally, they become narrower for horizons between 37- and 45-hour ahead, since predicted power is again at a low level. However, for the very last look-ahead times, one notices that intervals for nominal coverage rates greater than 80% have high upper bounds. This reflects the possibility of large negative prediction errors, even if such errors are unlikely.

5.8.1 Linear opinion pool vs. Adapted resampling

Two approaches for the combination of error distributions have been introduced, based on the linear opinion pool and adapted resampling methods. While the first one is based on the weighting of the probability distributions in a probabilistic sense, the second uses

the weights from the fuzzy inference model (5.36) for defining the share of each sample in the multi-sample resampling scheme. Our first aim is to compare the intervals resulting from these two approaches. For that purpose, we evaluate the quality of the interval forecasts produced from the point predictions given by M1, M2 and M3 for the two wind farms. In all the following evaluation exercises, the predictive distributions of wind generation will consist in a set of interval forecasts with nominal coverage rates ranging from 10 to 90%, with an increment of 10%. Regarding the set-up of both methods, the mapping of the forecast uncertainty is done by dividing the range of possible predicted power values into five zones, to which we associate triangular fuzzy sets. The size of the error samples is set to 300. Finally, we consider the case of 50 bootstrap replications for the adapted resampling approach.

Focus is given first to the reliability aspect, since we expect that the choice between the two approaches for combining probability distributions will mainly have an effect on the reliability of the resulting predictive distribution quantiles. Therefore, we estimate the actual coverage of the predictive distributions, and summarize this information in reliability diagrams that give the difference between the empirical and the nominal coverage rates, for the various estimated quantiles. Figures 5.5 and 5.6 depict the results for Tunø Knob and Klim respectively. These two diagrams are for the whole forecast length: displayed values correspond the average deviations over all the prediction horizons.

Consider in a first stage Figure 5.5 for explaining how to read such reliability diagrams and what kind of conclusions can be derived from their study. The x-axis gives the required probability, i.e. the nominal coverage rate of the predictive quantiles, and the various curves display the deviation (in %) from the 'perfect reliability' situation for which the empirical coverage of the quantiles would equal the nominal one. This ideal situation is represented by the dash-dot straight line. Then, a +1%-deviation for the quantile with nominal coverage rate 30% (for instance) actually tells that the empirical coverage estimated with Equation (5.73) is equal to 31%. Figures in the legend correspond to the average absolute deviation from the ideal case, over the range of nominal coverage rates (and also over the forecast length).

For the Tunø Knob case-study, the deviations from 'perfect reliability' are contained in a $\pm$3% envelope whatever the considered point prediction method or the interval estimation approach. The reliability of the intervals could be expected to be lower for low and high nominal coverage rates since it is harder to model the very tails of error distributions. This is not the case here. However, one notices a general trend, which is that quantiles for proportions below 0.5 are overestimated while quantiles above the median are underestimated. Prediction intervals are slightly too narrow on average. It should be understood here that having too narrow intervals is more likely than having too large intervals: methods for estimating future uncertainty usually rely on past experience of a given model performance and therefore do not integrate the additional uncertainty of predicting new data [15]. Average absolute deviations are between 0.86 and 1.23%, with slightly better results obtained from the application of the linear opinion pool approach. In a general manner, we conclude on an acceptable reliability of the probabilistic forecasts produced by both methods.

The Klim test case consists in a longer evaluation period (almost 19.000 series of two-

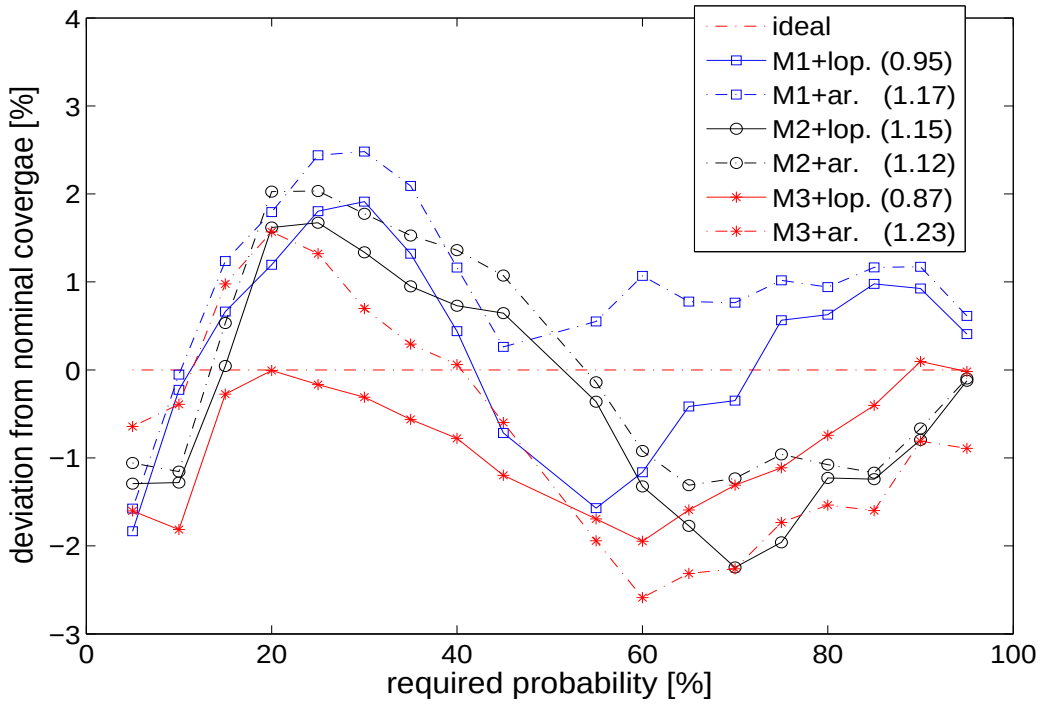

Figure 5.5: *Reliability diagrams for Tunø Knob. Results are given for the three point prediction methods and for the two implemented approaches (lop: linear opinion pool; ar: adapted resampling).*

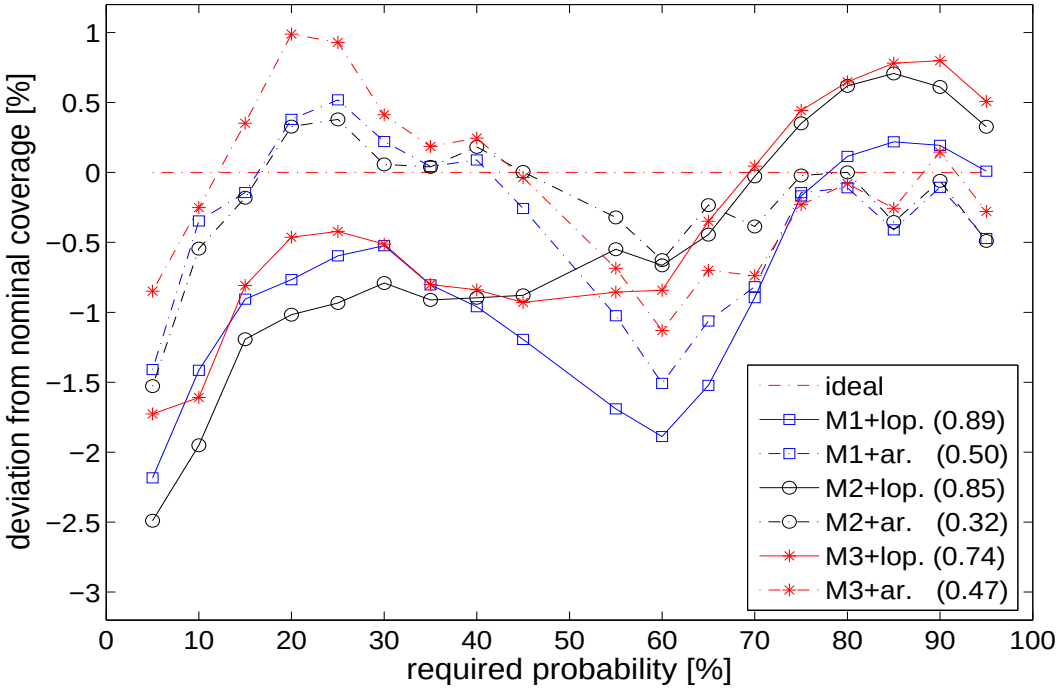

Figure 5.6: *Reliability diagrams for Klim. Results are given for the three point prediction methods and for the two implemented approaches (lop: linear opinion pool; ar: adapted resampling).*

days ahead forecasts) and can thus give more insight on the reliability of predictive distributions. In Figure 5.6, deviations from nominal coverage are in general lower than the ones witnessed when studying Tunø Knob. Average absolute deviations range from 0.32 to 0.89% only. These deviations are significantly lower for intervals estimated with the adapted resampling approach. There is a trend that the linear opinion pool quantiles underestimate the true quantiles (cf. left part of the reliability diagrams). The difference between the two approaches is less pronounced for quantiles with proportions higher than 0.5. Even if the calibration of adapted-resampling quantiles appears better than the one of linear-opinion-pool quantiles, we consider here also that both approaches yield reliable interval forecasts.

The second stage of the evaluation of the predictive distributions is carried out by using a scoring rule of the form of Equation (5.79) where the s_i functions are such that $s_i(p) = 4p$, $(i = 1, \ldots, 18)$, and $f(p) = -2p$ following Gneiting and Raftery [34], calculated as a function of the look-ahead time. The resulting score summarizes the skill of the predictive distributions described by the 18 quantiles estimated from both interval estimation methods. Given that we have accepted quantiles to be reliable (even if it is only a subjective result), the positively-oriented score can tell which method (and also which point prediction method used as input) leads to the 'best' predictive distributions.

Figure 5.7 gives the evolution of the skill score as a function of the horizon for Tunø Knob. Figures in the legend correspond to the average skill score values over the forecast length. The skill score steadily decreases as the look-ahead time augments. This meets the general statement that it is harder to predict for lead times further in the future, which was already discussed and illustrated for the case of point predictions of wind power in [64]. Also, we see that the skill scores of predictive distributions generated from the linear opinion pool and adapted resampling approaches are rather close: they actually coincide when point predictions are provided by the M3 method, though there are significant differences for the cases of M1 and M2. The average values shown in the legend tell that adapted-resampling predictive distributions are better than the ones resulting from the linear-opinion-pool combination approach.

The way the skill score evolves as a function of the look-ahead time for the Klim case-study is shown in Figure 5.8. Similarly, the skill score values of predictive distributions are rather close, with a slight advantage for the ones estimated with adapted resampling. But, an interesting point is that the choice of a given point prediction method as input has an influence on the quality of the resulting predictive distributions. Indeed, it appears that considering M1 leads to better probabilistic forecasts for Tunø Knob (but not over all the forecast length), whereas considering M2 is better for Klim. Note that since the predictive distributions are actually estimations of the error distributions related to point forecasts, a point prediction method with sharper error distributions will yield sharper probabilistic forecasts. Though, this comment is of course only valid if the prediction interval estimation approach has a real ability to reflect the error distribution associated to a given point forecast.

To conclude on that comparison of the two approaches for the probabilistic distribution combination problem, we can say that predictive distributions estimated from the adapted resampling approach prove to have a higher skill than the ones resulting from

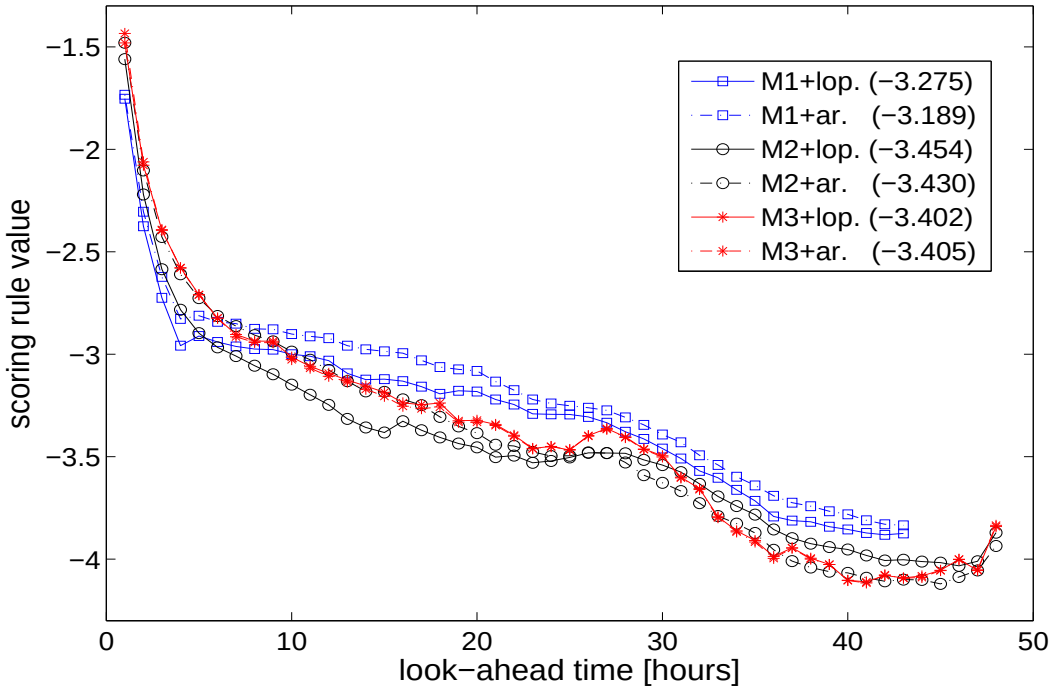

Figure 5.7: *Skill score as a function of the horizon for Tunø Knob. Results are given for the three point prediction methods and for the two implemented approaches (lop: linear opinion pool; ar: adapted resampling).*

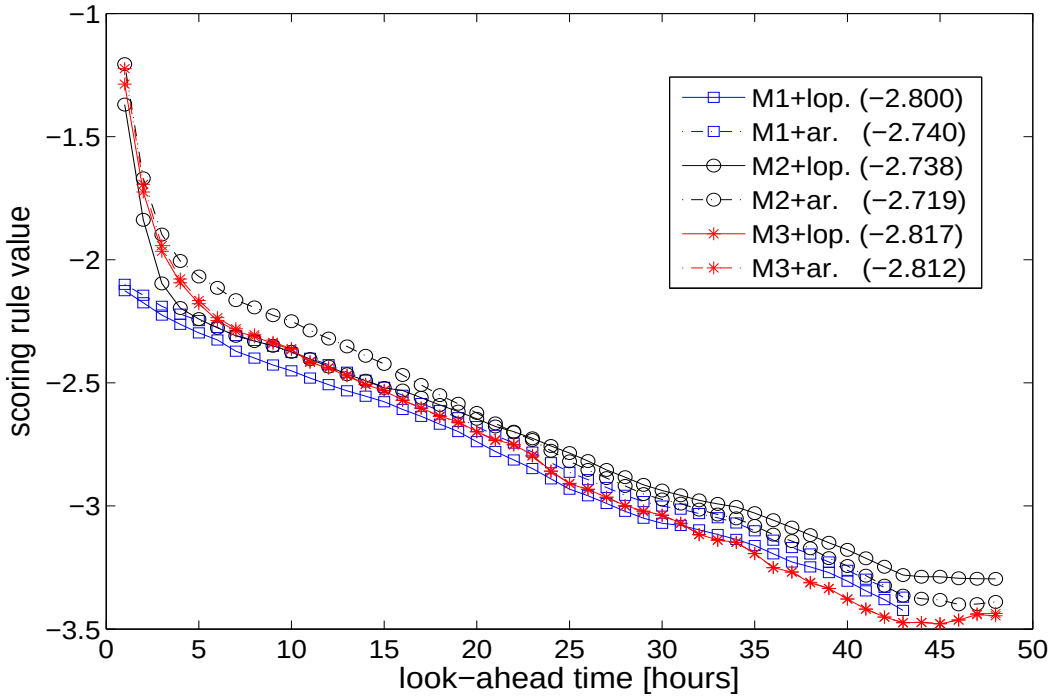

Figure 5.8: *Skill score as a function of the horizon for Klim. Results are given for the three point prediction methods and for the two implemented approaches (lop: linear opinion pool; ar: adapted resampling).*

the more classic linear opinion pool approach. On the case-study with the longer evaluation period, the reliability of adapted resampling quantiles is significantly higher. Also, for both case-studies and for the three point prediction methods considered as input, this method leads to higher values of the skill score, which encompasses all the aspects of the evaluation of probabilistic forecast quality. This is why we will focus on this approach in the following Paragraphs, and illustrate the influence of its degrees of freedom, on reliability, sharpness and resolution. This study comprises a sensitivity analysis of its performance and will result in general guidelines for its application to further case-studies or alternatively for online forecasting exercises.

5.8.2 Influence of the fuzzy mapping of the forecast conditions

The idea of introduced method is to propose a situation-dependent assessment of the forecast uncertainty: fuzzy logic is used for mapping several zones with different characteristics of the prediction error distributions. As explained previously, we concentrate here on the variation of the forecast uncertainty as a function of the level of predicted power. The range of possible predicted power values is divided into several ranges, to which we associate triangular fuzzy sets. It is expected that increasing the number of fuzzy sets will mainly have a positive effect on increasing the resolution of predictive distributions. In a general manner, increasing the resolution of probabilistic forecasting methods will augment their value for the management or the trading of wind generation (as long as they are still reliable). Three possibilities are envisaged: using only one fuzzy set on the power range (which is equivalent to using the classical Williams-Goodman empirical approach, cf. Paragraph 5.4.1), and a mapping with alternatively 3 or 5 fuzzy sets. We set the sample size to 300 elements and the number of bootstrap replications to 50. The considered case-study for that sensitivity analysis is Tunø Knob. The point predictions used as input are provided by the M2 method.

For assessing in a first stage the reliability of the probabilistic forecasts produced with these three settings, we use the reliability diagrams depicted in Figure 5.9. The deviations from nominal coverage are given as the average deviations over the whole range of look-ahead times. The figures given in the legend are the average absolute deviations from 'perfect reliability' (over the various nominal coverage rates and forecast horizons).

One sees from Figure 5.9 that the deviations from 'perfect reliability' are of the same order for the various settings: they are within $\pm 2.5\%$. Even if it is not the primary aim of the mapping, it seems that using several fuzzy sets permits to increase the overall reliability of estimated quantiles. In this example, the average absolute deviation is 1.40 % for Williams-Goodman intervals, whereas it is equal to respectively 1.05 and 1.12% for the two other settings with 3 and 5 fuzzy sets.

Then, we turn our attention to sharpness and resolution. Since sharpness proved to be similar for the three settings and that the power curve mapping is mainly expect to impact the resolution of predictive distributions, we focus here only on the latter quality aspect. We base our evaluation of the resolution property of the intervals on the σ-diagrams depicted in Figure 5.10, which give the standard deviation of the interval size as a function of the interval nominal coverage rate. As an example, we focus on

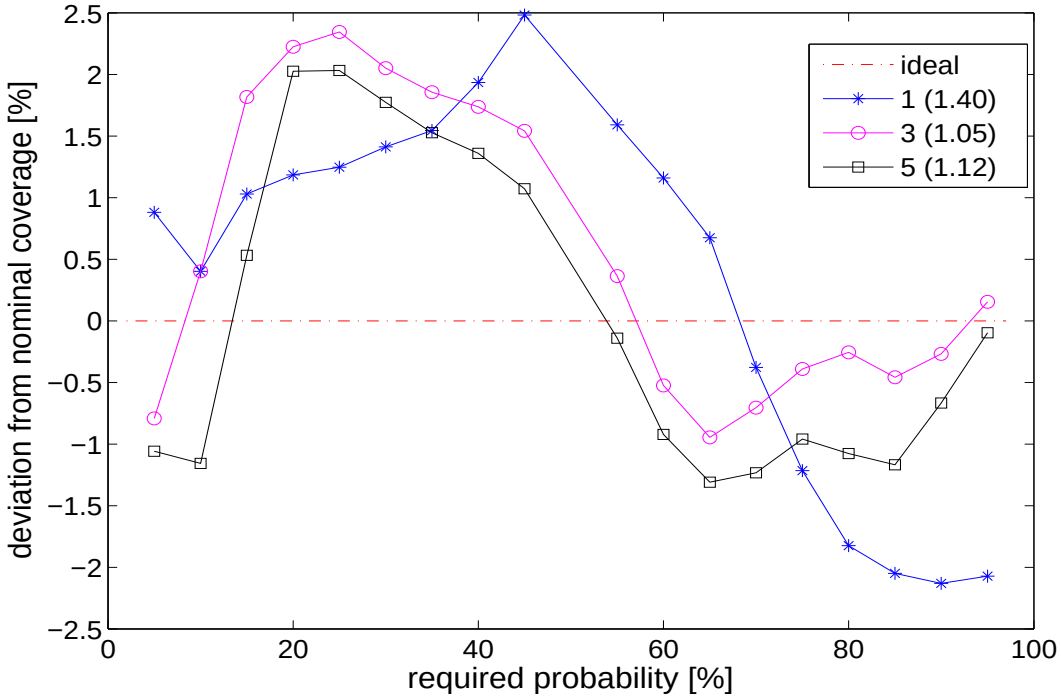

Figure 5.9: *Reliability diagrams for evaluating the influence of the power curve mapping on the resulting probabilistic forecasts' reliability.*

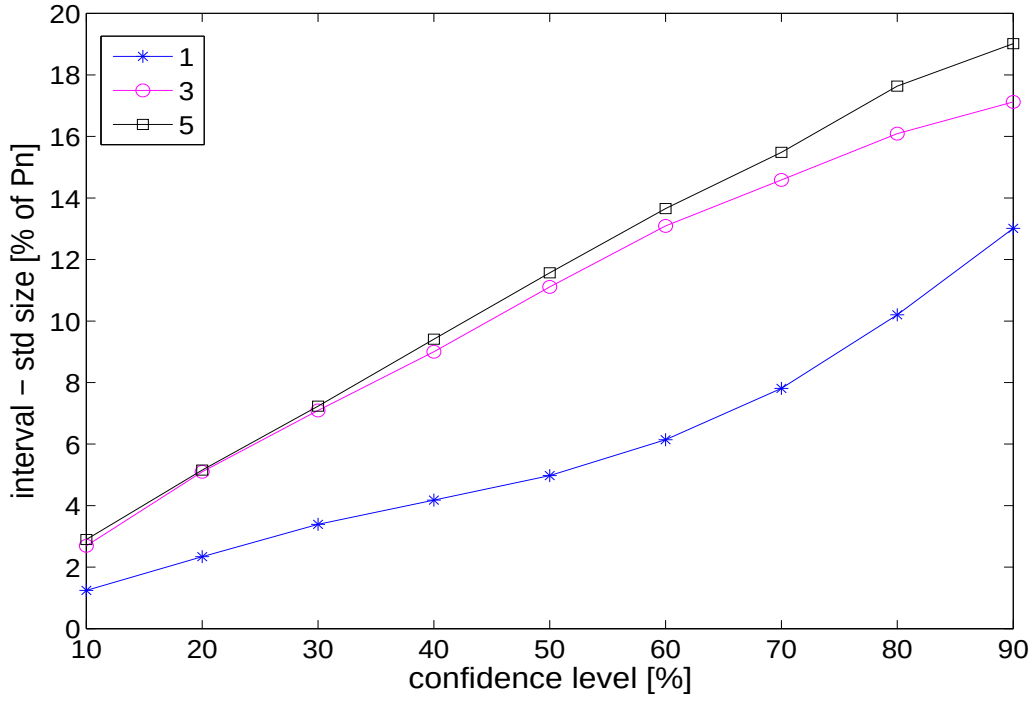

Figure 5.10: *σ-diagrams for 24-hour ahead forecasts for evaluating the influence of the power curve mapping on the resulting probabilistic forecasts resolution.*

the σ-diagrams of 24-hour ahead probabilistic forecasts, but similar conclusions could be derived if looking at σ-diagrams for 24-hour ahead probabilistic forecasts, but similar conclusions could be derived if looking at σ-diagrams for other look-ahead times. When going from Williams-Goodman to adapted resampling intervals, the resolution is significantly augmented, whatever the degree of confidence. For the example of the 50%-confidence prediction interval, the standard deviation of the interval size is actually multiplied by a factor 3. Also, one sees that by using more fuzzy sets for mapping the power curve, the resolution can be increased even more, mostly for high degrees of confidence. This means that the method has a better ability to differentiate the tail ends of predictive distributions. However, considering 3 or 5 subsets of forecast conditions leads to the constitution of the same number of error samples (respectively 3 and 5 samples of prediction errors). Therefore increasing the method resolution has a cost, which is the time needed for filling the error samples. Here a minimum of respectively 900 and 1500 series of point predictions are necessary for filling the samples. Note that this is does not appear to be a restriction for the application of the method, since intervals can be estimated even if all the samples are not full. The only consequence is that predictive distributions may not be as reliable as it would be expected in a first period of the application. This is certainly a reason why predictive distributions produced with the 5-fuzzy-set configuration are slightly less reliable than the ones from the 3-fuzzy-set configuration here. Finally, even if we have focused on the resampling method, we have noticed that the fuzzy mapping of the forecast conditions has a similar influence on the skill of the linear opinion pool approach.

5.8.3 Influence of the sample size

The second part of the study concerns the influence of the sample size, i.e. the number of past prediction errors, on the skill of the estimated intervals provided by the adapted resampling method. Intuitively, considering more past errors should permit to better understand the uncertainty of the process and thus to augment the reliability of estimated predictive distributions. However, relying on very large error samples would make the method less dynamic. Here, the number of fuzzy sets is set to five and the number of bootstrap replications to 50. We produce probabilistic forecasts with error samples containing 50, 100, 200 and 300 elements.

The reliability diagrams displayed in Figure 5.11 show how the sample size affects the predictive distributions' reliability. The absolute average deviation from 'perfect reliability' greatly diminishes as we use more elements in the adapted resampling procedure. This absolute average deviation is divided by 2 if considering the last 300 errors instead of dealing with the last 50 only. Also, one notes from the reliability diagrams that the trend to have too narrow intervals (i.e. overestimated quantiles if below the median and underestimated if above the median) diminishes when the sample size is increased.

For evaluating the sharpness of the interval forecasts, we superpose δ-diagrams for the various method settings (Figure 5.12). This Figure is for 24-hour ahead probabilistic forecasts. In a general manner, the average interval size ranges from $\sim$5% of nominal power for intervals at a 10% degree of confidence to $\sim$50% for those associated with a

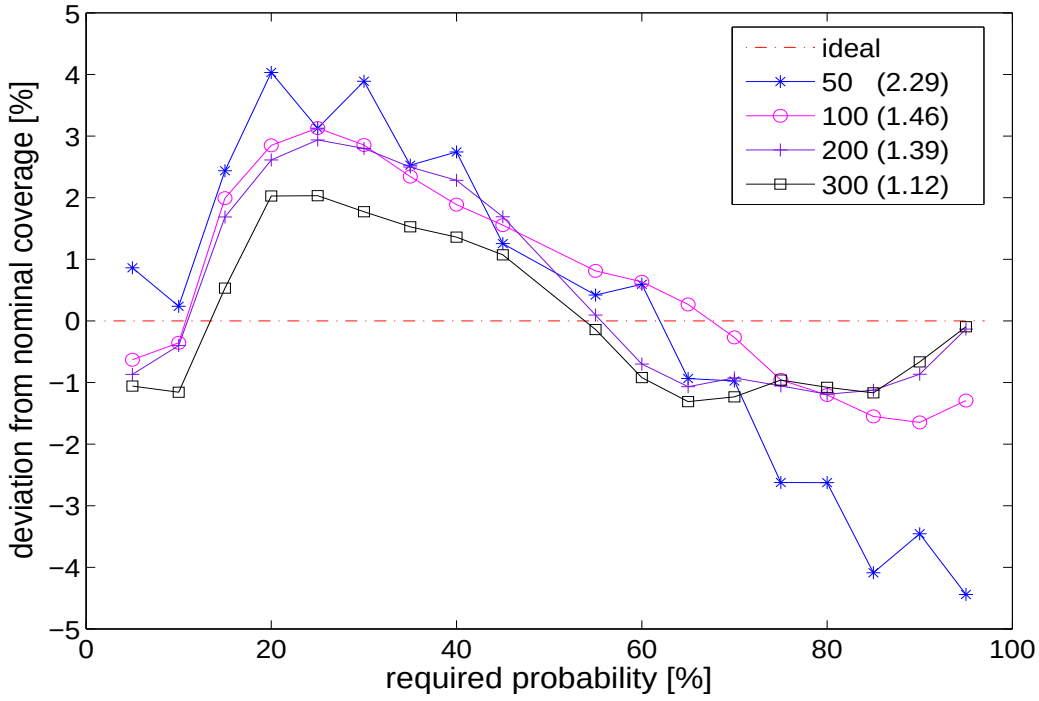

Figure 5.11: *Reliability diagrams for evaluating the influence of error sample size on the resulting probabilistic forecasts' reliability.*

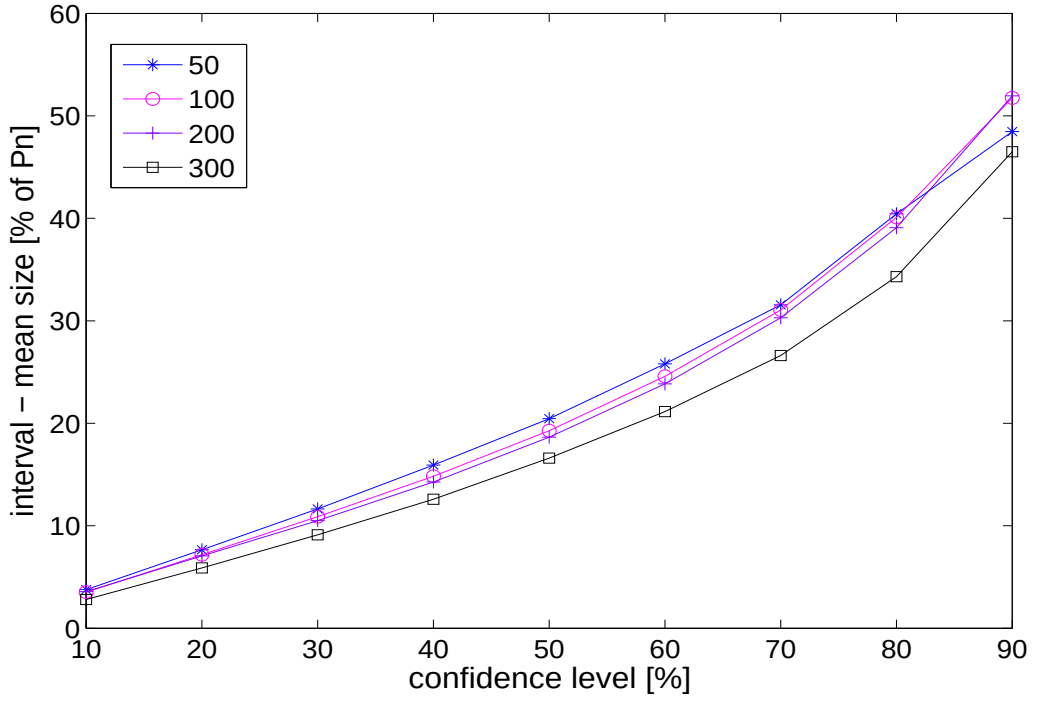

Figure 5.12: *δ-diagrams for 24-hour ahead forecasts for evaluating the influence of the error sample size on the resulting probabilistic forecasts' sharpness.*

90% degree of confidence. Whatever the nominal coverage rate, the average size decreases when considering more past prediction errors for estimating predictive distributions. This diminution in the mean size is up to 10% when going from 50 to 300 sample elements. Therefore, by increasing the sample size, we improve both the forecasts reliability and their sharpness. That parameter does not have a significant effect on the resolution of predictive distributions.

This study of the influence of the sample size also tells what can be the expected quality of probabilistic forecasts in a first period of an online forecasting exercise, as error samples are filled when new point predictions are provided. At the very beginning, error samples are empty and the approach we have developed cannot be used for estimating interval forecasts. However, one understands by looking at Figures 5.11 and 5.12 that an acceptable performance is already attained with a minimum number of 50 elements. Therefore, defining a larger sample size does not affect the quality of estimated interval in that sample-filling period. Note that one may also envisage to extract error samples from an offline forecasting exercise with the wind farm of interest, and use these samples for initializing the prediction interval estimation method for the online application.

5.8.4 Influence of the number of bootstrap replications

In the last part of the present sensitivity analysis, we turn our attention to the influence of the number of bootstrap replications on the quality of resulting predictive distributions of wind generation. For the adapted resampling method, the number of bootstrap replications correspond to the number of combined error samples created by following the fuzzy inference model (5.36). This degree of freedom is not present in the linear opinion pool approach. Augmenting the number of bootstrap replications then translates to considering more alternative scenarios for estimating the quantiles of predictive distributions. For better illustrating the influence of that parameter, we focus on smaller samples or errors. Here, that sample size is set to 50. The range of possible predicted power values is still mapped with 5 triangular fuzzy sets.

Primarily, we concentrate on the way the reliability of predictive distributions evolves with the number of bootstrap replications. Figure 5.13 gathers the reliability diagrams of these distributions when estimated after 1, 10, 100 and 1000 resampling steps. Again, average absolute deviations from 'perfect reliability' are given in the legend. Results are again for the whole forecast length. One notices that reliability is significantly increased when augmenting the number of bootstrap replications, up to 100 replications. However, it seems that increasing that parameter value is not necessary, since reliability remains at a similar level. Figure 5.14 then depicts the evolution of the skill score with the forecast horizon, where the skill score is defined in the same manner than in Paragraph 5.8.1. The curves for the various number of bootstrap replications are quite close, but one sees from the average values in the legend that the score values augment when considering more resampling steps. Better reliability contributes to augmenting the overall quality of predictive distributions. And, by considering 1000 resampling steps this overall quality is even slightly higher, certainly because probabilistic forecasts get sharper. But, as resampling methods are CPU-demanding, it is not desirable to use more and more boot-

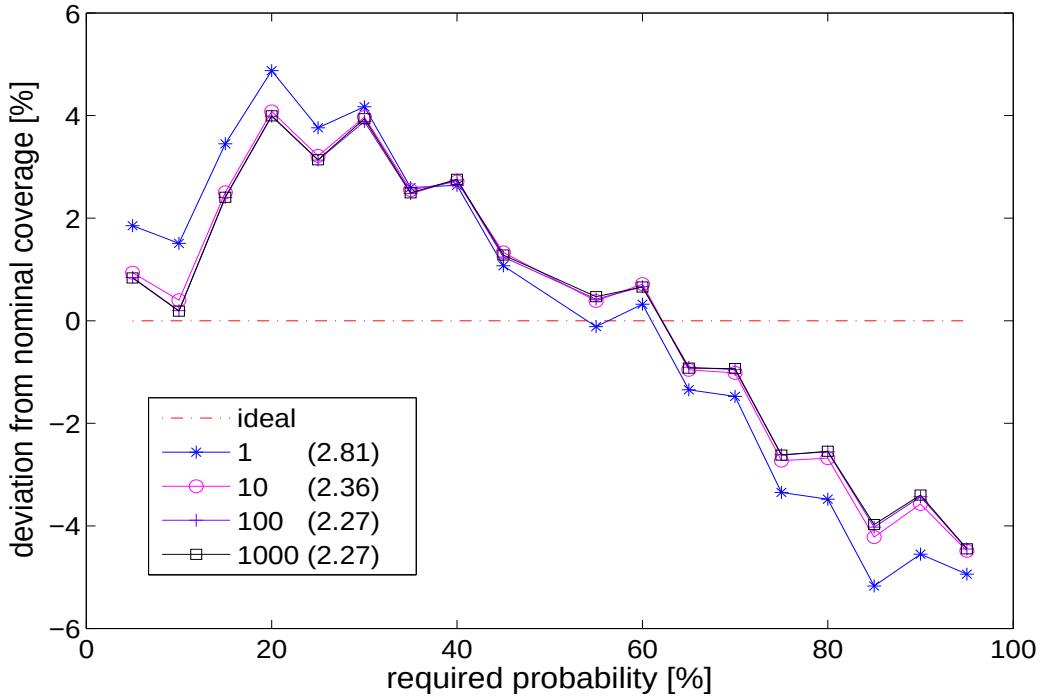

Figure 5.13: Reliability diagrams for evaluating the influence of the number of bootstrap replications on the resulting probabilistic forecasts' reliability.

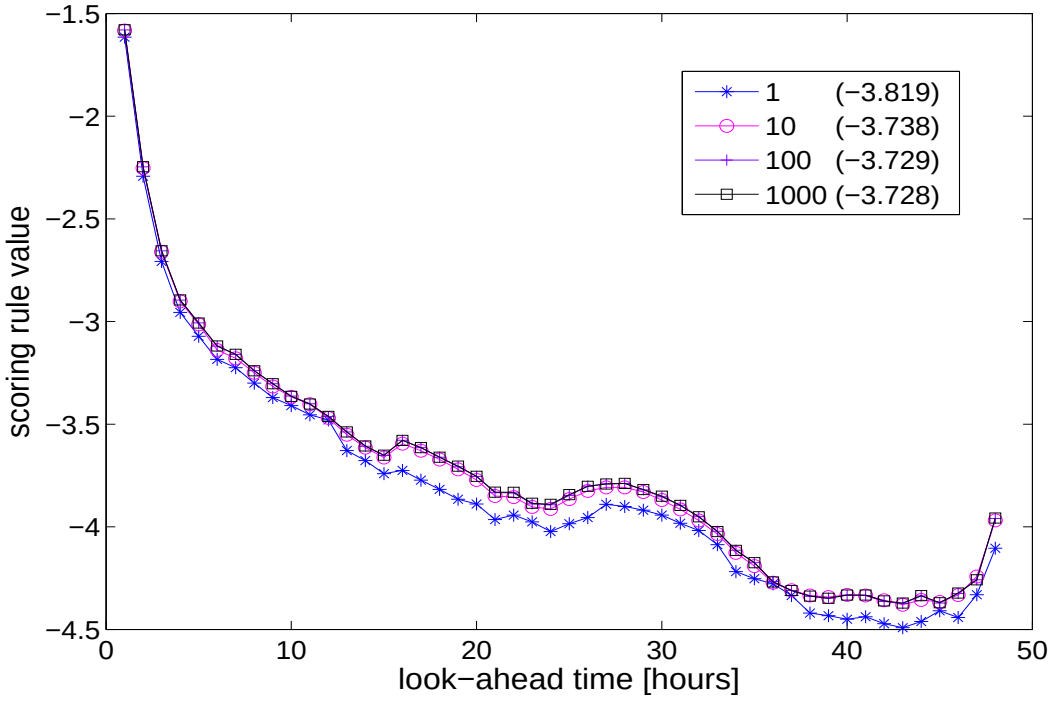

Figure 5.14: Influence of the number of bootstrap replications on the overall quality of predictive distributions. Overall quality is assessed with the skill score, given as a function of the forecast horizon.

strap replications if not necessary. Therefore, in the frame of an online application, one has to find a trade-off between the desired quality and the time that may be needed for computing probabilistic forecasts.

5.9 Conclusions

A generic method for assessing online the uncertainty of short-term wind power forecasts in the form of prediction intervals has been introduced. The developed method is designed for the case of nonstationary, nonlinear and bounded time-series of prediction errors. It has a non-parametric and empirical nature, since it considers recent prediction errors for estimating predictive distributions of wind generation. A great advantage of that method is that it permits to construct the whole distribution of errors at once, and can thus be used for estimating several prediction intervals without needing several models. Also, by having this empirical view, we encompass all the sources of uncertainty that may come from the input data, the chosen prediction model and its parameters, etc. However, the proposed method requires a subject-matter expertise of the process of interest, since the mapping of the forecast conditions related to various characteristics of the prediction error distributions has to be carried out by the analyst. Here, that expertise follows from the study of the characteristics of the state-of-the-art point prediction methods described in Anemos Deliverable Report 2.1 [64]. A fuzzy inference model is proposed, which permits to produce conditional error distributions given the forecast conditions, in the form of combined probability distributions. Predictive distributions of wind generation are then obtained by dressing point predictions with related conditional distributions of forecast errors. Two approaches for the combination of probability distributions resulting from the fuzzy inference model have been described. On the one hand, we applied the so-called linear opinion pool, which is a classical method in the probability combination literature. On the other hand, we have introduced an original approach referred to as adapted resampling. Such a method follows from the basic idea of resampling methods, which consists in thinking that more information can be extracted from a sample of data by cleverly going through that sample a certain number of times. In our case, the multi-sample resampling scheme is used for estimating the quantiles of the combined error distributions from samples representing the individual ones.

We have thoroughly demonstrated the quality of this method for the estimation of prediction intervals of wind power by evaluating its statistical performance. For that purpose, we have gathered a set of relevant skill scores, measures and diagrams, in a non-parametric framework suitable for assessing the developed method's properties. These properties include the reliability, sharpness and resolution of interval forecasts. The verification framework allowed us to conclude on the superiority of the adapted resampling method on the linear opinion pool approach. Also, we have considered some of the criteria for illustrating the influence of the method parameters on the various required properties for prediction intervals: mapping the forecast conditions increase the resolution of the resulting probabilistic predictions, while augmenting the sample size or the number of bootstrap replications mainly has an effect on the reliability property. Finally, we have given guidelines regarding the method configuration if applied for online fore-

casting exercises, since it definitely has an operational nature.

It was clearly shown that the approach is suitable for application to current state-of-the-art point prediction methods of wind generation. Our opinion is that the quality (more precisely the sharpness) of predictive distributions produced from such methods is bounded by the quality of the point predictions used as input. This follows from the fact that here probabilistic forecasts are based on the modeling of the point predictor's error distributions. The sharper these distributions, the sharper the resulting probabilistic forecasts. Therefore, further research works should go towards direct probabilistic forecasting of wind generation, in order to verify if by releasing the constraint of using power point predictions as input one can further increase the quality of predictive distributions.

Chapter 6

General Conclusions

In this report, we have described the methods developed in the frame of the Anemos project for estimating the uncertainty of point predictions of wind power production.

Since a part of the prediction error directly comes from the meteorological forecasts used as input, a possibility for telling on the confidence one may have in the wind power predictions has been to identify weather situations related to different levels of prediction error. It has been shown there were significant differences in average prediction accuracy among the different identified weather classes. Further focusing on the meteorological aspects for providing skill forecasts will be an important direction for future developments.

Another part of the work has consisted in using a model for the conversion of wind speed prediction errors to power prediction errors (with the local derivative of an explicit power curve) for associating point predictions with error bands. Such error bands inform on the magnitude of likely deviations from the predictions. Alternatively, two non-parametric methods have been described for estimating quantiles of predictive distributions of wind power output. By estimating several quantiles with different proportions, one may associate point forecasts with prediction intervals or even with full predictive distributions of expected generation at a given lead time. The two methods that have been introduced are based on quantile regression and on an empirical-type approach integrating an expert model. The quality of their output have been evaluated on several case-studies. Both methods are promising for online application.

Regarding future research works, it will be of particular interest to further develop on methods for probabilistic forecasting of wind power output, which could be either parametric or non-parametric. Ensemble predictions of meteorological variables can also be considered as input for that purpose. And, in order to compare the various methods that exist in the literature today, as well as the ones that are going to be developped in the next few years, it will be of particular importance to enrich and generalise the verification framework introduced in the present report. Verification frameworks are the basis for justifying new developments and evaluating their related benefits.

Probabilistic predictions are provided here for each look-ahead time, and do not adress the issue of the correlation in forecast errors. Such correlation exists within prediction series (that is, for consecutive look-ahead times), and also for successive prediction series. This is due to the fact that forecast errors typically occur if the prevailing weather

situation deviates from the predicted one, and this has a time scale of several hours. Proposing models for the correlation of these errors, or integrating this correlation information in probabilistic prediction models will significantly improve the value of wind power forecasts.

Finally, a direction for further research is to relate the quality of probabilistic predictions (that is, in terms of their statistical performance) and their value for the end-users, which will be measured by their increased benefits resulting from the use of such advanced forecasting methodologies. For that, we will have to consider different decision-making processes, either for the management of power systems integrating a significant share of wind power, or for the participation of wind power producers in electricity markets.

Appendix A

Parametric additive quantile models in "R"

The "quantreg" library or package is not part of a standard "R" installation. To install the package start "R" and issue the command:

install.packages("quantreg")

However, please read the help-page before issuing this command.

Assume that the data is contained in a data frame named **train** with columns **y**, **x1**, and **x2**. Furthermore, data for which quantile forecasts must be computed is assumed to be contained in a similar data frame named **test** with columns **x1** and **x2**.

The R-script shown in Table A.1 illustrates how parametric additive quantile models can be fitted, how the results can be visualized, and how forecasts can be produced.

Periodic bases [21] can be constructed from the output of the function **bs**. This is done by the S-PLUS/R function **pb.bse** found in the file **periodic.bases.q** at `http://www.imm.dtu.dk/~han/pub` which has been used in this paper. The restriction that the function approximated by the periodic basis integrates to zero over the period is imposed on the periodic basis using the function downloadable as **bint0.q**.

Table A.1: *R-script. Functions and operators are indicated by the bold font, comments start with "## ".*

```
## Load required libraries:
library(splines)
library(quantreg)

## Make natural spline bases with 10 columns and knots placed
## according to the quantiles of x1 and x2:
basis.x1 <- ns(train$x1, df = 10, intercept = F)
basis.x2 <- ns(train$x2, df = 10, intercept = F)

## Fit 25% and 75% quantile models (1 denotes the intercept):
fit25 <- rq(train$y ~ 1 + basis.x1 + basis.x2, tau = 0.25)
fit75 <- rq(train$y ~ 1 + basis.x1 + basis.x2, tau = 0.75)

## Estimated coefficients:
coef(fit25)
coef(fit75)

## Plot of fitted values of the estimates related to x1:
intercept.avg <- (coef(fit25)["(Intercept)"] + coef(fit75)["(Intercept)"])/2
matplot(train$x1,
        cbind(basis.x1 %*% coef(fit25)[grep("basis\\.x1", names(coef(fit25)))]
              + coef(fit25)["(Intercept)"] - intercept.avg,
              basis.x1 %*% coef(fit75)[grep("basis\\.x1", names(coef(fit75)))]
              + coef(fit75)["(Intercept)"] - intercept.avg
        ))

## Plot of fitted values of the estimates related to x2:
matplot(train$x2,
        cbind(basis.x2 %*% coef(fit25)[grep("basis\\.x2", names(coef(fit25)))]
              + coef(fit25)["(Intercept)"] - intercept.avg,
              basis.x2 %*% coef(fit75)[grep("basis\\.x2", names(coef(fit75)))]
              + coef(fit75)["(Intercept)"] - intercept.avg
        ))

## Forecast for data frame 'test':
test.b.x1 <- ns(test$x1,
                knots = attributes(basis.x1)$knots,
                Boundary.knots = attributes(basis.x1)$Boundary.knots,
                intercept = attributes(basis.x1)$intercept)
test.b.x2 <- ns(test$x2,
                knots = attributes(basis.x2)$knots,
                Boundary.knots = attributes(basis.x2)$Boundary.knots,
                intercept = attributes(basis.x2)$intercept)
##
## Comment: It is very important that the bases for prediction are
## constructed independently from the test data, i.e. by supplying the
## knots etc. as outlined above.
##
qForecast <- data.frame(Q25 = cbind(1, test.b.x1, test.b.x2) %*% coef(fit25),
                        Q75 = cbind(1, test.b.x1, test.b.x2) %*% coef(fit75))
## Print forecasts:
qForecast
```

Appendix B

Implementation of a Module for Online Estimation of Prediction Intervals of Wind Generation

The methods for the estimation of prediction intervals of wind generation have been developed for online operation. In the frame of the thesis, we have also developed a module (in C++ programming language) which is integrated in the ANEMOS prediction platform. In the present Appendix, we present the main characteristics of this module, from its configuration by the analyst to the visualization of prediction intervals, via some details about its operation in an online environment.

Module setup

In the ANEMOS prediction platform, the necessary information for the operation of the different prediction modules are stored in a Static Data Repository (SDR), which is the database with all the static data. These information include the characteristics of the considered wind farm (e.g. geographical coordinates), of the NPWs (e.g. temporal resolution), etc. Then, each of the prediction modules may utilize a local configuration file in which are stored its specific parameters. The configuration file for the interval forecasting module contains the following parameters:

***Method*:** defines the chosen approach for the computation of prediction intervals, i.e. *linear opinion pool* or *adapted resampling*,

***File path*:** defines the path to the directory where are stored the local memory files necessary for the operation of the module (i.e. history of predictions and influential variables, and error samples),

***Number of intervals*:** defines the number of prediction intervals to be computed. Whatever the chosen approach, several intervals can be computed at once since they model the whole predictive distribution of wind generation for every horizon,

Algorithm B.1: *The chain of tasks to be carried out by the uncertainty estimation module at each prediction time.*

step 1.	Load the module configuration file and the relevant parameters from the SDR
step 2.	Retrieve the new power measures, power predictions and influential variable values from the TSDR
step 3.	Load the memory files containing stored values of power predictions and influential variables
step 4.	Load the memory files containing the error samples
step 5.	Calculate the prediction errors from collected power measures and stored predictions
step 6.	Determine the forecast conditions related to calculated prediction errors
step 7.	Update and save the files containing power predictions and influential variables
step 8.	Update error samples given the forecast conditions, and save them in the memory files
step 9.	Use the fuzzy inference model for determining the distributions of prediction errors associated to every power predictions
step 10.	Apply either the linear opinion pool or the adapted resampling method for estimating the bounds of the prediction intervals with the required nominal coverage rates
step 11.	Save the prediction intervals in the TSDR

Nominal coverage rate (i)**:** defines the nominal coverage rate of the i^{th} prediction interval to be computed. Such value is comprised between 0 and 100%,

Sample size**:** defines the size of the samples of past prediction errors,

Resampling times**:** defines the number of bootstrap replications if one chooses to utilize the adapted resampling approach,

Number of influential variables**:** defines the number of influential variables,

Influential variable (i)**:** gives the type of the i^{th} influential variable, such as predicted wind power or forecast wind speed for instance,

Number of ranges (i)**:** defines the number of ranges of values that have to be considered for the i^{th} influential variable. For instance, if this parameter is set to 5 for the predicted power variable, then the range of possible predicted power values is divided into 5 ranges. To each of these ranges is associated a triangular fuzzy set,

[low,up] (i,j)**:** defines the lower and upper bound of the j^{th} range of values for the i^{th} influential variable. The fuzzy set related to this range of values is defined accordingly.

One sees from this list of parameters that the configuration of the module can be

tailored to the considered application, depending on the analyst's expertise on the specificities of that application.

Operation

The module provides interval forecasts with the same forecast length and resolution than the point prediction module it is associated to. In most of the cases, predictions are produced for look-ahead times up to 48-hour ahead, with an hourly resolution. Also, prediction intervals are provided with the same frequency of update than the point prediction module it is associated to. In general, forecasts are updated on an hourly basis for statistical methods and only when NWPs are provided for the case of physical methods (e.g. every 6 hours when considering HIRLAM meteorological predictions as input).

A scheduler is at the heart of the ANEMOS prediction platform. It manages the retrieval of onsite measures and meteorological forecasts, as well as the operation of the various modules. Since interval forecasts are associated to series of point predictions, the module for uncertainty estimation is run just after the point forecasting one. The chain of tasks to be carried out by the uncertainty estimation module is similar to the chain described by Algorithm 4.1, with some additional steps dedicated to the communication with the Time-Series Data Repository (TSDR), which is the database containing all the dynamic data, as well as file management issues. This chain is given by Algorithm B.1.

During online operation, power measurements, power predictions and influential variables values (e.g. wind speed forecasts) may be erroneous or missing. Hence, the second step of the above Algorithm, which consists in the retrieval of these data, also integrates a data checking procedure. If power measurements are missing, the prediction errors cannot be calculated and thus error samples are not updated. And, if power predictions or influential variable values are missing or erroneous, interval forecasts are not computed. Instead, series of "-99" values are returned (in step 11), as well as a message indicating that interval computation was not possible.

Results and Visualization

The ANEMOS platform is also composed by a man-machine interface. Such an interface allows the end-user to visualize historic power production, power predictions, as well as associated prediction intervals. Also, the end-user may use that interface for consulting reports on the performance of the various prediction methods over a given period.

Figure B.1 shows a general view of the man-machine interface. In the upper window is represented the island of Crete, with the 12 wind farms for which wind generation is predicted. For wind farms that are equipped with SCADA systems and thus for which power measurements are regularly stored in the TSDR, the interval forecasts produced from the previously described uncertainty estimation module can be visualized at the same time than the point predictions. The two lower windows of Figure B.1 display 48-ahead point predictions to which are associated prediction intervals with a nominal

coverage rate of 80%.

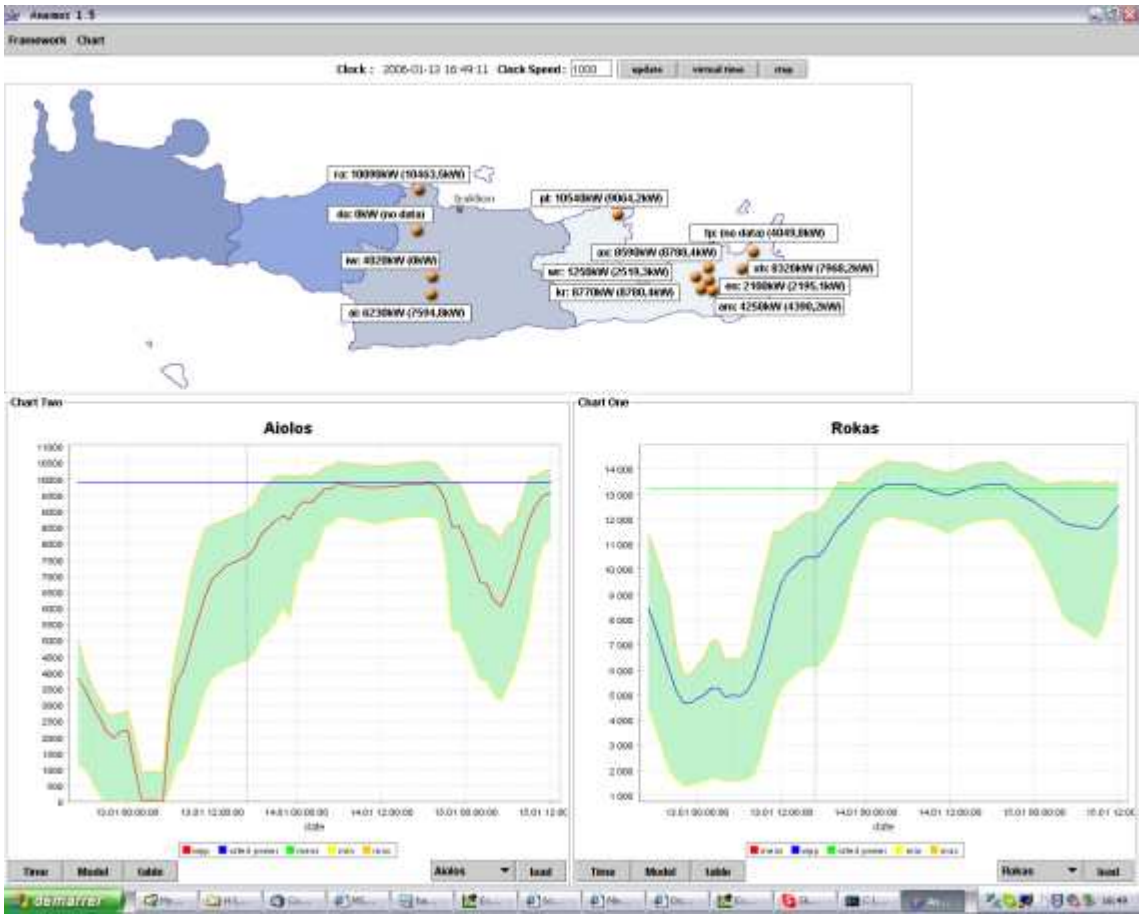

Figure B.1: *The man-machine interface of the ANEMOS prediction platform for a Windows XP operating system.*

Bibliography

[1] A. P. Alves da Silva and L. S. Moulin. Confidence intervals for neural network based short-term load forecasting. *IEEE Trans. on Power Syst.*, 15(4), November 2000.

[2] F. Atger. The skill of ensemble prediction systems. *Mon. Wea. Rev.*, 127:1941–1957, September 1999.

[3] R. T. Baillie and T. Bollerslev. Prediction in dynamic models with time-dependent conditional variances. *J. Econometrics*, 51:91–113, 1992.

[4] R.G. Barry and A.H. Perry. *Synoptic Climatology*. Methuen & Co Ltd, 1973.

[5] G. N. Bathurst, J. Weatherhill, and G. Strbac. Trading wind generation in short-term energy markets. *IEEE Trans. on Power Syst.*, 17(3):782–789, Aug. 2002.

[6] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.

[7] S. Bordignon and F. Lisi. Interval prediction for chaotic time-series. *Metron*, 59(3-4):117–140, 2001.

[8] G. E. P. Box and G. M. Jenkis. *Time-Series Analysis, Forecasting an Control*. Holden-Day, 1970.

[9] J. B. Bremnes. Probabilistic wind power forecasts by means of statistical model. In *Proceedings of the IEA R&D Wind Annex XI Joint Action Symposium on Wind Forecasting Techniques*, pages 49–58, Norrköping, Sweden, December 3–4 2002.

[10] J. B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7(1):47–54, January-March 2004.

[11] D.S. Broomhead and G.P. King. *On the qualitative analysis of experimental dynamical systems*, chapter Nonlinear Phenomena and Chaos. Adam Hilger, 1986.

[12] B. G. Brown, R. W. Watz, and A. H. Murphy. Time-series models to simulate and forecast wind speed and wind power. *J. App. Met.*, 23(8):1184–1195, 1984.

[13] J. G. Carney, P. Cunningham, and U Bhagwan. Confidence and prediction intervals for neural network ensembles. In *Proc. of the International Joint Conference on Neural Networks 1999*, 1999. paper 2090.

[14] W. S. Chan, S. H. Cheung, and K. H. Wu. Multiple forecasts with autoregressive time-series models: case-studies. *Mathematics and Computers in Simulation*, 64:421–430, 2004.

[15] C. Chatfied. *Time-Series Forecasting*. Chapman & Hall/CRC, 2000.

[16] P. F. Christoffersen. Evaluating interval forecasts. *Intern. Econom. Rev.*, 39(4):841–862, 1998.

[17] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203, 1999.

[18] M. P. Clements. *Evaluating Econometric Forecasts of Economic and Financial Values*. Palgrave Macmillan, 2005.

[19] M. P. Clements and N. Taylor. Bootstrapping prediction intervals for autoregressive models. *Int. J. Forecasting*, 17(2):247–267, 2001.

[20] W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

[21] C. de Boor. *A Practical Guide to Splines*. Springer Verlag, Berlin, 1978.

[22] R. De Veaux, J. Schumi, J. Schweinsberg, D. Shellington, and L. H. Ungar. Prediction intervals for neural networks via nonlinear regression. *J. Amer. Stat. Ass.*, 40(4):273–282, November 1998.

[23] R. Doherty and M. O'Malley. A new approach to quantify reserve demand in systems with significant installed wind capacity. *IEEE Trans. on Power Syst.*, 20(2):587–595, May 2005.

[24] R. Dybowski and S. J. Roberts. Confidence intervals and prediction intervals for feed-forward neural networks. In R. Dybowski and V. Gant, editors, *Clinical Applications of Artificial Neural Networks*. Cambridge University Press, 2000.

[25] B. Efron. Bootstrap methods: another look at the jackknike. *Ann. Statist.*, 7:1–26, 1979.

[26] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[27] B. Everitt. *Cluster Analysis*. Heinemann Educational Books, 1974.

[28] A. Fabbri, T. G. Gómez San Román, J. R. Rivier Abbad, and V. H. Méndez Quezada. Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Trans. on Power Syst.*, 20(3):1440–1446, 2005.

[29] U. Focken. *Leistungsvorhersage räumlich verteilter Windkraftanlagen unter besonderer Berücksichtigung der thermischen Schichtung der Atmosphäre*. PhD thesis, University Carl von Ossietzky, Oldenburg, Germany, 2003.

[30] C. Genest and K. J. McConway. Allocating the weights in the linear opinion pool. *J. Forecasting*, 9:53–73, 1990.

[31] F.-W. Gerstengarbe and P.C. Werner. Katalog der Grosswetterlagen Europas (1881-1998) nach Hess und Brezowsky. Potsdam-Institut für Klimafolgenforschung, Potsdam, Offenbach a.M., 1999.

[32] G. Giebel, R. Brownsword, and G. Kariniotakis. State of the art on short-term wind power prediction. ANEMOS Deliverable Report D1.1, available online: http://anemos.cma.fr, June 2003.

[33] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. Technical report, University of Washington, Department of Statistics, May 2005. Technical report no. 483.

[34] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. Technical report, University of Washington, Department of Statistics, 2004. Technical report no. 463.

[35] B. Grassberger. Do climatic attractors exist? *Nature*, 323:609–611, 1986.

[36] M. Grigoletto. Bootstrapping prediction intervals for autoregressions: some alternatives. *Int. J. Forecasting*, 14(4):447–456, 1998.

[37] G. J. Hahn and W. Q. Meeker. *Statistical Intervals - A Guide for Practitioners*. John Wiley & Sons, 1991.

[38] P. Hall and A. Rieck. Improving coverage accuracy of nonparametric prediction intervals. *J. Royal Stat. Soc.*, 63(4):717–725, 2001.

[39] T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, 129, 2000. Notes and Correspondence.

[40] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, London/New York, 1990.

[41] T. Heskes. Practical confidence and prediction intervals. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*. MIT Press, Cambridge, 1997.

[42] D. Hou, E. Kalnay, and K.K. Droegemeier. Objective verification of the SAMEX '98 ensemble forecast. *Mon. Wea. Rev.*, 129:73–91, 2001.

[43] J. J. G. Hwang and A. A. Ding. Prediction intervals for artificial neural networks. *J. Amer. Stat. Ass.*, 92:748–757, 1997.

[44] R. J. Hyndman. Highest-density forecast regions for non-linear and non-normal time-series models. *J. Forecasting*, 14:431–441, 1995.

[45] M. Jørgensen and D. I. K. Sjøberg. An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. *Information & Software Technology*, 45:123–136, 2003.

[46] L.S. Kalkstein, G. Tan, and J.A. Skindlov. An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Climate Appl. Meteor.*, 26:717, 1987.

[47] H. Kantz and T. Schreiber. *Nonlinear Time-series Analysis*. Cambridge University Press, 1997.

[48] A. B. Koehler. An inappropriate prediction interval. *Int. J. Forecasting*, 6(4):557–558, 1990.

[49] R. W. Koenker and G. W. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

[50] R. W. Koenker and V. D'Orey. [Algorithm AS 229] Computing regression quantiles. *Applied Statistics*, 36:383–393, 1987.

[51] S. Kotz, N. L. Johnson, and C. B. Read, editors. *Encyclopedia of statistical sciences*, volume 5. John Wiley & Sons, 1982.

[52] J.-P. Lam and M. R. Veall. Bootstrap prediction intervals for single period regression forecasts. *Int. J. Forecasting*, 18(1):125–130, 2002.

[53] L. Landberg. Short-term prediction of local wind conditions. Risø-R-702(EN), Risø National Laboratory, 1994.

[54] L. Landberg, G. Giebel, L. Myllerup, J. Badger, H. Madsen, and T.S.Ñielsen. Poor-man's ensemble forecasting for error estimation. In *Proc. of the 2002 AWEA WindPower Conference, Portland, Oregon (USA)*, 2002.

[55] M. Lange. *Analysis of the uncertainty of wind power predictions*. PhD thesis, University Carl von Ossietzky, Oldenburg, Germany, 2003.

[56] M. Lange. On the uncertainty of wind power predictions - Analysis of the forecast accuracy and statistical distribution of errors. *Trans. of the ASME, J. Solar Energy Eng.*, 127(2):177–194, May 2005.

[57] M. Lange and U. Focken. *Physical Approach to Short-term Wind Power Prediction*. Springer Verlag, 2005. ISBN 3-540-25662-8.

[58] M. Lange and D. Heinemann. Accuracy of short term wind power predictions depending on meteorological conditions. In *CD-Proc. of the 2002 Global Windpower Conference, Paris, France*, April 2002.

[59] M. Lange and D. Heinemann. Relating the uncertainty of short-term wind speed predictions to meteorological situations with methods from synoptic climatology. In *Proceedings of the European Wind Energy Conference & Exhibition*, Madrid, Spain, 2003. http://www.ewea.org.

[60] E.N. Lorenz. A new approach to linear filtering and prediction problems. *Deterministic nonperiodic flow*, 20:130–141, 1963.

[61] A. Luig, S. Bofinger, and H. G. Beyer. Analysis of confidence intervals for the prediction of regional wind power output. In *Proc. of the 2001 European Wind Energy Conference, EWEC'01, Copenhagen, Denmark*, pages 725–728, June 2001.

[62] H. Madsen, P. Pinson, H. Aa. Nielsen, T. S. Nielsen, and G. Kariniotakis. Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering*, 2006. In Press.

[63] S. Makridakis, M. Hibon, E. Lusk, and M. Belhadjali. Confidence intervals - An empirical investigation of the series in the M-competitions. *Int. J. Forecasting*, 3(3-4):489–508, 1987.

[64] I. Marti and co authors. ??? ANEMOS Deliverable Report D2.1, available online: http://anemos.cma.fr, -??- 2004.

[65] J. Mason. Definition of technical terms in forecast verification and examples of forecast verification scores. In *Proc. of the IRI Workshop on Forecast Quality, New York*, October 2000.

[66] S. McNees. Forecast uncertainty: can it be measured? mimeo, Federal Reserve Bank of Boston, Boston, Massachusetts (USA), 1995.

[67] K. Mönnich. *Vorhersage der Leistungsabgabe netzeinspeisender Windkraftanlagen zur Unterstützung der Kraftwerkseinsatzplanung*. PhD thesis, University Carl von Ossietzky, Oldenburg, Germany, 2000.

[68] A. H. Murphy. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, 8(2):281–293, 1993.

[69] C. Nicholis and G. Nicholis. Is there a climatic attractor? *Nature*, 321:529–532, 1984.

[70] H. Aa. Nielsen. Analysis and simulation of prediction errors for wind power productions reported to nord pool. IMM-Eltra project report, Informatics and Mathematical Modeling, Technical University of Denmark, Denmark, 2002. in Danish.

[71] H. Aa. Nielsen and H. Madsen. Analyse og simulering af prædiktionsfejl for vindenergiproduktion ved indmelding til NordPool. Technical report, Informatik og Matematisk Modellering, Danmarks Tekniske Universitet, Lyngby, 2002. I samarbejde med Eltra a.m.b.a.

[72] H. Aa. Nielsen, T. S. Nielsen, and H. Madsen. Using meteorological forecasts for short term wind power forecasting. In *Proceedings of the IEA R&D Wind Annex XI Joint Action Symposium on Wind Forecasting Techniques*, pages 49–58, Norrköping, Sweden, December 3–4 2002.

[73] H. Aa. Nielsen, T. S. Nielsen, and H. Madsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. In *Proc. of the 2004 European Wind Energy Conference, EWEC'04, Scientific Track, London, United Kingdom*, pages 34–38, November 2004.

[74] H. Aa. Nielsen, T. S. Nielsen, H. Madsen, and K. Sattler. Wind power ensemble forecasting. In *CD-Proc. of the 2004 Global Windpower Conference, Chicago, Illinois (USA)*, March 2004.

[75] T. S. Nielsen, H. Aa. Nielsen, and H. Madsen. Prediction of wind power using time-varying coefficient-functions. In *Proceedings of the XV IFAC World Congress*, Barcelona, 2002.

[76] T.S. Nielsen, H. Madsen, and H.S. Christensen. WPPT – a tool for wind power prediction. In *Proceedings of the Wind Power for the 21st Century Conference*, Kassel, Germany, 2000.

[77] T.N. Palmer. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, 63:71–116, 2000.

[78] P. Pinson and G. Kariniotakis. Wind power forecasting using fuzzy neural networks enhanced with on-line prediction risk assessment. In *CD-Proc. of the 2003 Bologna Power Tech Conference, IEEE Power Tech 2003, Bologna, Italy*, June 2003.

[79] P. Pinson and G. Kariniotakis. On-line adaptation of confidence intervals based on weather stability for wind power forecasting. In *CD-Proc. of the 2004 Global Windpower Conference, Chicago, Illinois (USA)*, March 2004.

[80] P. Pinson and G. Kariniotakis. On-line assessment of prediction risk for wind power production forecasts. *Wind Energy*, 7(2):119–132, May-June 2004.

[81] N. Ravishankar, L. Shiao-Yen Wu, and J. Glaz. Multiple prediction intervals for time-series: comparison of simultaneous and marginal intervals. *J. Forecasting*, 10:445–463, 1991.

[82] J. J. Reeves. Bootstrap prediction intervals for ARCH models. *Int. J. Forecasting*, 20(2):237–248, 2004.

[83] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, 130:1653–1660, June 2002. Notes and Correspondence.

[84] B. H. Sass, N. W. Nielsen, J. U. Jørgensen, B Amstrup, M. Kmit, and K. S. Mogensen. The operational DMI-HIRLAM system 2002-version. Technical Report 02-05, Danish Meteorological Institute, 2002. http://www.dmi.dk.

[85] R. Schrodin. Quarterly report of the operational nwp-models of the deutscher wetterdienst. vol. 20, Deutscher Wetterdienst, Offenbach a.M., 1999., 1999.

[86] M. Shahgedanova, T.P. Burt, and T.D. Davis. Synoptic climatology of air pollution in Moscow. *Theor. Appl. Climatol.*, 61:85, 1998.

[87] T. R. Stewart. Uncertainty, judgement, and error in prediction. In D. Sarewitz, R. A. Pielke, and R. Byerly, editors, *Prediction: Science, Decision Making, and the Future of Nature*, pages 41–57. Island Press, 2000.

[88] M. Stone. The opinion pool. *Annals of Mathematical statistics*, 32:1339–1342, 1961.

[89] J. W. Taylor and D. W. Bunn. Investigating improvements in the accuracy of prediction intervals for combination of forecasts: a simulation study. *Int. J. Forecasting*, 15(3):325–339, 1999.

[90] Z. Toth, O. Tallagrand, G. Candille, and Y. Zhu. Probability and ensemble forecasts. In I.T. Jolliffe and D.B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Ltd, 2003.

[91] L.-W. Wang. *Adaptive fuzzy systems and control*. Prentice Hall, 1994.

[92] W. H. Williams and M. L. Goodman. A simple method for the construction of empirical confidence limits for economic forecasts. *J. Amer. Stat. Ass.*, 66(336):752–754, 1971. Applications Section.

[93] R. L. Winkler. A decision-theoretic approach to interval estimation. *J. Amer. Stat. Ass.*, 67:187–191, 1972.

[94] B. Yarnal. *Synoptic climatology in environmental analysis - A primer*. Belhaven Press, 1993.