SafeWind

Collaborative project funded by the European Commission under the 7th Framework Program, Theme 2007-2.3.2: Energy

"Multi-scale data assimilation, advanced wind modelling & forecasting with emphasis to extreme weather situations for a safe large-scale wind power integration"

Grant Agreement N°: 213740

DELIVERABLE DP-6.6

"Predictability measure and communication of information on uncertainty to forecast users"

DOCUMENT TYPE	Deliverable
DOCUMENT NAME:	swind.deliverable_Dp-6.6.doc
VERSION:	V1.1(*)
DATE:	2012.028.14
CLASSIFICATION:	R0: General public
STATUS:	Approved

Abstract:
Wind power is a renewable energy source that is rapidly growing worldwide. Since wind speed fluctuates in time and space, the power output of wind turbines has large variations. The integration of the generated power into the electricity grid gives rise to several challenges, especially when high wind power penetration levels are reached. System operators now recognize that wind power forecasting up to two to three days ahead contributes to a secure and economic power system operation. Apart from point forecasts of the wind farm output for the coming hours and days, of major importance is the development of tools for the online assessment of the uncertainty of these forecasts and the provision of information about the risk for large forecast errors.

Probabilistic forecasts of wind generation have received significant attention in the academic literature but have yet to be translated into best practice within industry. State-of-the-art methods have the potential to provide end-users with a situation-dependent estimation of forecast uncertainty, thereby facilitating improved decision-making. The primary concern of many end-users receiving forecasts relates to knowing in advance whether or not the forecasts will be accurate. Improving our ability to estimate the tails of the forecast density, that is the performance of probabilistic forecasts in providing reliable information about potential extremes and associated large forecast errors, is at the heart of this challenge.

For this reason, SafeWind has focused extensively on developing methods for constructing, evaluating and communicating probabilistic forecasts of wind power generation, with a particular emphasis on the tails of predictive distributions. The underlying processes that govern wind power generation are influenced by several meteorological variables and modelling is complicated by the nonlinearity of the transformation between wind speed and wind power. An array of new approaches for probabilistic forecasting have now been developed within the SafeWind consortium for addressing these particular characteristics.

In the first part of this report, we present an overview of the approaches available for assessing predictability of extremes, accounting for uncertainty and communicating this information in a coherent format to end-users. Describing uncertainties relating to extremes is not trivial and this challenge will be explored throughout the report. In contrast to extreme value analysis that offers a static assessment of extremes, the approach developed here provides a dynamical assessment of risk, which is directly relevant to the challenges now facing the wind energy sector. Wind power forecast ensembles, derived from ensembles of numerical weather predictions offer a means of assessing predictability. A practical example of communicating the spatial profile of risk levels is given for the particular extreme events associated with the storm Xynthia. Both the spread of the ensemble forecast and the absolute forecast error are considered as a means of communicating forecast uncertainty.

The ensemble members represent different scenarios and the spread among the members is expected to provide a measure of predictability. In the second part of this report, the spread is captured by risk indices with the purpose to give information about the expected forecast error, particularly the risk for large errors. As a starting point, a previously proposed definition of a risk index, based on the weighted standard deviation of the ensemble members, is evaluated. The evaluation is made on three French wind farms on different temporal and spatial scales. Furthermore, a number of alternative definitions of risk indices are proposed and examined to assess whether they could be more suitable. Finally an investigation of how risk indices can be presented to end-users is performed, including suggestions and evaluations of a number of alternative methods.

Results show that the average forecast error increases as the ensemble spread increases. The ensemble spread is shown to be well captured by risk indices, which are useful in providing information about the uncertainty of the forecasts and give an indication of the risk for large forecast errors. The choice of prediction model, particularly with regard to how the ensemble spread is affected when meteorological ensembles are converted to power ensembles, is found to be important. Although the alternative risk indices do not always outperform the previously used definition, some of them give similar performance while being easier to calculate and interpret. Finally, it is found that the way in which risk indices are presented to end-users is important. Some of the proposed alternative methods are found to outperform those used in previous studies.

AUTHORS[1], REVIEWERS			
MAIN AUTHOR/EDITOR:	P. McSharry		
AFFILIATION:	University of Oxford		
ADDRESS:	Smith School of Enterprise and the Environment, 75 George Street, Oxford OX1 2BQ, UK.		
TEL.:	+44 1865 614943		
EMAIL:	patrick.mcsharry@smithschool.ox.ac.uk		
FURTHER AUTHORS:	L. von Bremen (OLDENBERG), Erik Holmgren, Nils Siebert, Georgios Kariniotakis, Robin Girard (ARMINES)		
PEER REVIEWERS:			
REVIEW APPROVAL:	Approved :		Rejected (improve as indicated below) :
SUGGESTED IMPROVEMENTS:	*For a long list of remarks make reference to another document*		
APPROVER:			

VERSION HISTORY			
VERSION[2]:	DATE:	COMMENTS, CHANGES, STATUS:	PERSON(S):
1.1	14.08.2012	Combination of Oxford, Oldenberg and ARIMINES reports	P.E. McSharry

STATUS, CONFIDENTIALITY, ACCESSIBILITY									
STATUS:				CONFIDENTIALITY:				ACCESSIBILITY:	
S0	Approved/Released	X		**R0**	General public	**X**		Private web site	X
S1	Reviewed			**R1**	Restricted to project members			Public web site	
S2	Pending for review			**R2**	Restricted to European Commission			Paper copy	
S3	Draft for comments			**R3**	Restricted to WP members + PL				
S4	Under preparation			**R4**	Restricted to Task members +WPL+PL				

PL: *Project leader* **WPL:** *Work package leader* **TL:** *Task leader*

[1] **The authors of this document are solely responsible for its content, which does not represent the opinion of the European Community**

[2] **VERSION NAMING : V0.x** draft before peer-review approval, **V1.0** at the approval, **V1.x** minor revisions, **V2.0** major revision

Part 1

Contents

1. Introduction

1.1 Connecting modellers and end-users

There is a substantial gap between the motivations, activities and incentives of modellers producing forecasts and end-users who are tasked with the challenge of making time-critical decisions based on the outputs of the forecasts and a host of other sets of information and constraints. In the context of SafeWind, we consider modellers who apply their techniques to identify an appropriate quantitative model (supported by physical, mathematical and statistical expertise) for producing forecasts of wind power generation at the level of an individual wind farm, region or nation. As far as the modeller is concerned, the resulting forecast performance may be quantified using an appropriate statistical score which depends on the format of the forecast. For example, scores such as the mean absolute error, check function or continuous ranked probability score (CRPS) may be utilised to evaluate point, quantile and density forecasts respectively. Recommendations for best practice in forecast evaluation is given in Dp6.2 entitled "Methodology for the evaluation of probabilistic forecasts" (McSharry et al., 2009). In contrast, the decision-maker will have a specific utility function, usually not transparent to the decision-maker herself and rarely conveyed to the modeller who may have been commissioned to construct the forecasts in the first place. This utility function, which may vary with time, will encapsulate the full range of rewards and penalties associated with forecasts that both under- and over-estimate the actual level of wind power forecasting.

1.2 Decision-making

Constructing a suitable framework for communicating forecasts and their inherent uncertainties hinges on access to the utility function of the decision-maker. Indeed it is possible that the utility function of the decision-maker is at odds with that of their organisation and with the requirements of the regulator. In any case, for the purpose of this report, we will focus on facilitating the specific decision-making at the level of an individual. The forecasts may therefore be viewed as a decision-support tool for this particular individual and the goal is to convey these forecasts in a format that is most convenient for helping that decision-maker to optimise their objectives in as far as they can be translated into the impacts of potential forecast errors.

1.3 Structure of the report

In the following, we offer a risk classification scheme, which can be deployed to communicate the results of any probabilistic forecasting approach to end-users. Section 2 describes the varied objectives from the point of view of the end-user that requires forecasts for improving their ability to make decisions. It discusses the specific challenges pertaining to the integration of wind power energy and focuses on the effect of uncertainty and the need for probabilistic forecasting of wind power. Section 3 describes the fluctuations observed in typical wind power time series and links these to the objectives in the previous section. Extreme events are at the heart of the problem and we discuss how an early warning system is required to facilitate timely responses to such events. We present a methodology for providing an appropriate risk classification scheme and discuss its use in the wind energy sector. Section 4 gives a practical example whereby the uncertainty in spatial forecasts, measured using both the ensemble spread and the absolute forecast error, is communicated using spatial maps. Section 6 concludes and summarises the key points of the report.

2. Challenges, uncertainty and probabilistic forecasts

2.1 Objectives

The objectives of SafeWind in terms of communication of uncertainty information to end-uses in the field of wind energy can be broadly summarised as:

- To develop new mathematical and statistical algorithms appropriate for adaptive modelling and analysis of streaming wind power data
- To provide a measure of the risk of ramps
- To identify of novel or anomalous behaviour that might indicate future ramp events
- To provide end-users with a risk index that quantifies the likelihood of variability that warrants special attention

2.2 Wind power generation

The variable nature of wind generation poses a number of challenges that act as a limiting factor for wind penetration. The variability of wind power output creates energy imbalances that need to be resolved by the system operator with the allocation of spinning reserve and induces penalties for wind generators that reflect over- or under- production in the real-time market. In this context, accurate short- to medium-term (minutes to one week ahead) forecasts of wind power output are crucial both for technical and financial issues. These include the smooth operation of the network, the economic dispatch of reserve capacity, the minimisation of risk for wind generators, the scheduling of maintenance operations, the impact on the price of electricity, and, in the longer-term, the evaluation of wind energy resource allocation and the design of wind farms. A common theme in all these issues is the need to manage risk.

Wind power time series from individual wind farms in the form of metered data are collected and summed to calculate the aggregate wind power production in different countries. For convenience, when constructing statistical models, this aggregate wind power is usually presented as a normalised value, reflecting the fraction of the total installed capacity at each instant in time. The time series demonstrates underlying non-stationary processes driven by spatiotemporal wind fields and the changes in the diversity of wind farms that are online at any moment in time. A range of statistical time series models have been developed and compared using the aggregate wind power generated in the Republic of Ireland (Lau and McSharry, 2010).

2.3 Forecast uncertainty

The need to cope with forecast uncertainty is receiving increasing attention from the scientific community at present. Unfortunately modellers have a tendency to play down uncertainty and to be overconfident when explaining the predictive power of their models. Furthermore, many end-users view uncertainty as a weakness in the model or perhaps an inadequacy in the particular scientist that constructed the model. Due to the nonlinear dynamics of the atmosphere and the possible chaotic processes therein, weather forecasts will always be uncertain and the level of this uncertainty grows with the forecast horizon. For this reason, a scientist that explicitly communicates the level of uncertainty in their forecast should be applauded. Judging by the reaction of the BBC to the failure of the UK Met Office's seasonal forecasts (McSharry, 2010a), it may still be some time before probabilistic forecasts are either accepted or made available for public consumption (McSharry, 2010b).

While it is relatively straightforward to produce forecasts, it is much more difficult to evaluate and compare forecasts from a range of competing service providers. McSharry et al. (2009) describe a methodology for evaluating both point and probabilistic forecasts of wind power generation.

Throughout the SafeWind project, there has been a strong emphasis on using appropriate performance metrics, known as proper statistical scores, for evaluating probabilistic forecasts. This is a fundamental step in ensuring that the best methods for generating forecasts are identified and made available to the scientific community. An appropriate method for evaluating the reliability of probabilistic forecasts has been developed in SafeWind (Pinson et al., 2010).

2.4 Variability indices

The common complaint about wind power is that is highly variable. Providing a means of measuring the degree of this variability is a starting point for many analyses. Davy et al. (2010) defined a variability index for wind speed time series by filtering out the low frequency component and then taking a moving average. Their variability index was defined as the standard deviation of a band-limited signal in a moving window, and depends on four parameters: the order of the filter (integer greater than one), the upper and lower frequencies of the extracted signal, and the width of the moving window. Anastasiades and McSharry (2012) propose a parsimonious variability index for wind power time series, which depends on only two parameters. The original wind power time series is first smoothed using a moving average with a window of size m. Finally the new variability index is given by calculating the standard deviation of the smoothed signal using a window of size n. In addition, Anastasiades and McSharry (2012) propose three alternative measures of variability, which employ the interquartile range, 5% and 95% quantiles respectively. Examples of the four variability indices are shown in Figure 1.

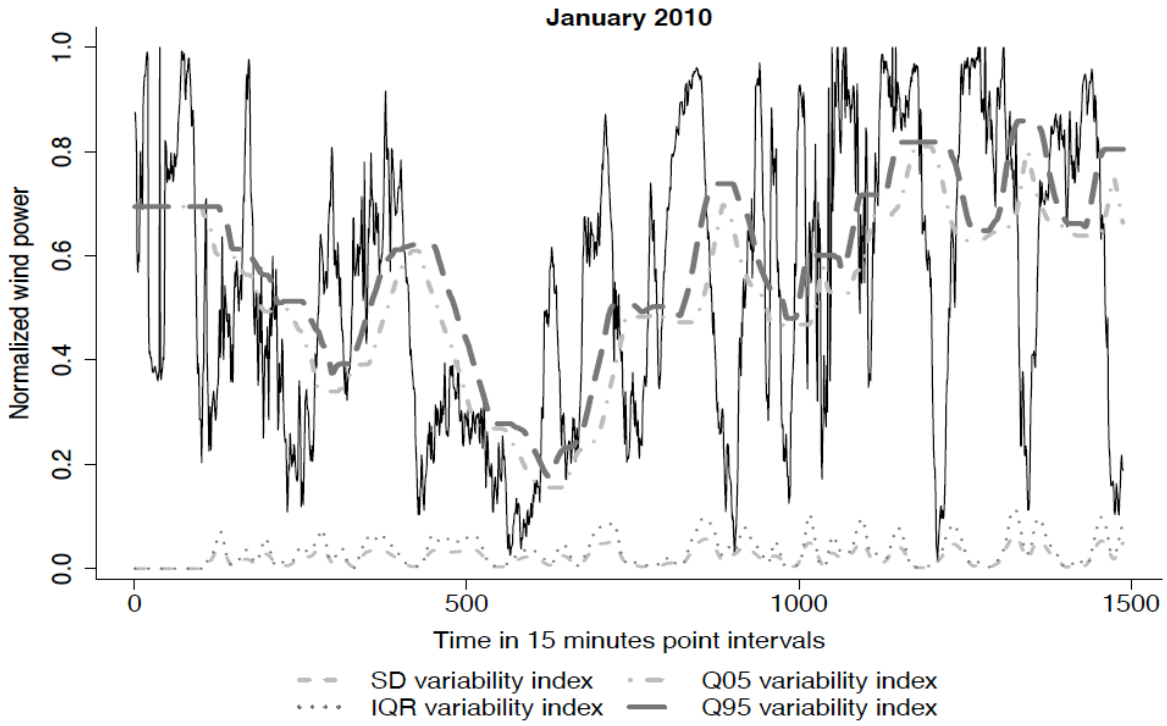


Figure 1: Wind power time series during one month and the four variability indices.

3. Extreme events and early warning systems

3.1 Extreme events

Extreme fluctuations in the amount of wind power production are of great importance in the wind energy sector and offer considerable challenges to the transmission system operators (TSOs). These changes are known as “ramps” in the industry and both decreases and increases must be managed carefully. Downward ramps are caused by declines in the wind speed but can also result from increases in the wind speed above the turbine cut-off threshold (typically 25 m/s) due to the need for certain turbines to be shut down in order to protect them from windstorm damage. A sudden decline in the amount of wind power requires additional capacity to be brought online in order to make up the shortfall and the electricity generators that can function in these short time frames are usually dependent on fossil fuels. Unexpected rises in wind power generation may also lead to problems

since the network may not be able to cope with the additional flows due to transmission constraints and may therefore require curtailment of the wind energy.

A useful means of communicating the variability in wind power fluctuations is to calculate the cumulative probability of negative and positive changes respectively. The size of the fluctuations varies with the time scale and we illustrate changes over 15 minutes, 30 minutes, one hour and four hours in Figure 2. The probability of ramps over 15 minutes and a corresponding normal distribution with the mean and standard deviation estimated empirically is shown in Figure 3. This shows that the probability of both negative and positive extreme events is greater than that predicted by the normal distribution. This has important implications for risk management. In particular the assumption that the normal distribution is appropriate would lead to a substantial underestimate of the risk.

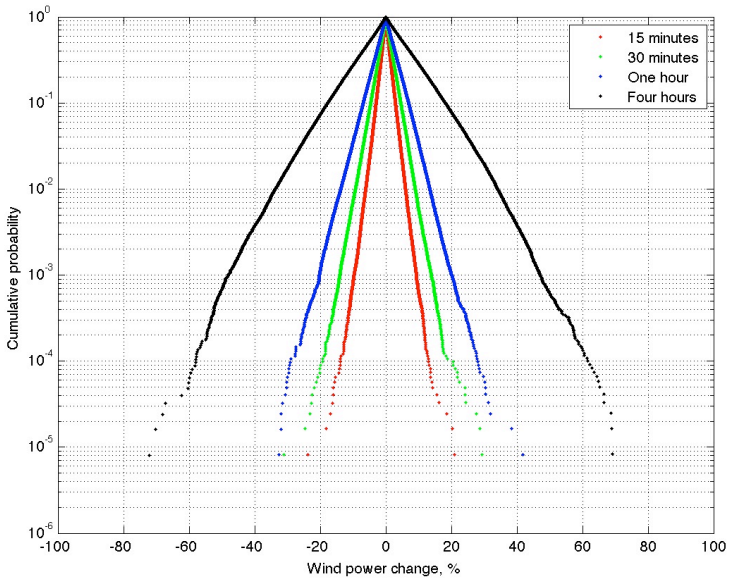


Figure 2: Cumulative probability of extreme fluctuations in wind power generation over horizons of 15 minutes, 30 minutes, one hour and four hours.

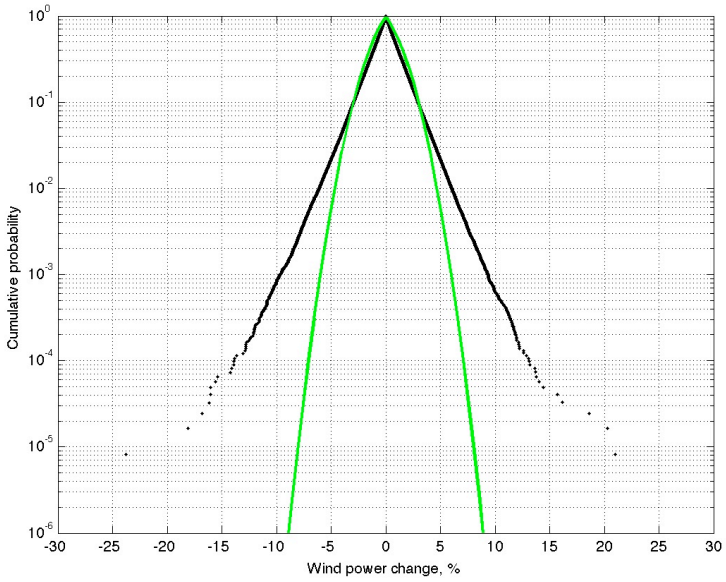


Figure 3: Cumulative probability of ramps over 15 minutes compared with the best fit provided by a normal distribution.

3.2 Early warning systems

The TSOs would benefit from a robust approach to providing warnings of forthcoming extreme events such as ramps, cut-offs and potential curtailment. The SafeWind project has made considerable progress in categorising, identifying and forecasting extreme events.

While classification typically relies on a binary decision process, we also wish to extend our analyses to a risk probability index, which might be utilised to construct a more sophisticated early warning system. For example, one might use the colour and risk levels currently employed by the Met Office for describing extreme weather events. We have adapted this approach for a generic early warning system as shown in **Table 1**. This has the advantage of mapping the risk level into four different categories and being able to provide specific advice for the action that should be taken in response to each.

In order to evaluate the performance of a risk forecasting system, we need to ascertain the specific utility function of the end-users or stakeholders that will benefit from the successful operation of the early warning system. This is not a straightforward procedure and relies on undertaking a systematic cost-benefit analysis of responding to alerts. Costs will be incurred whenever the authority reacts to a warning and this must be balanced with the potential savings made when acting to prevent or mitigate against an extreme event.

For example, we consider a simple method of labelling which rewards the early warning system for providing an alert during the period of ten days before the extreme event. The risk level rises linearly during this period indicating the likelihood of the forthcoming extreme event. **Figure 4** shows a schematic example of the labels, reflecting the idealised risk levels that may be employed to design such an early warning system. The increasing risk level that is shown as one approaches the extreme event is generally representative of what might be expected from an appropriate system. If an extreme event can be anticipated, it is likely that the ability to do so will increase with time, reaching a peak at the actual time of the event.

Table 1: Colours and risk levels that might be used for an operational early warning system. This gives details of the risk, headline, impact and advice in the form of specific actions that should be taken. Adapted from UK Met Office.

	Green	Yellow	Amber	Red	
Warning	None	Advisory	Advisory	Early	Flash
Risk	Very low <20%	Low ≥20% <40%	Moderate ≥40% <60%	High ≥60% <80%	Very high ≥80%
Headline	No extreme event expected	Low risk of extreme event	Moderate risk of extreme event	High risk of extreme event	Extreme event is imminent or occurring
Impact		Low risk of major damage to infrastructure	Moderate risk of major damage to infrastructure	High risk of major damage to infrastructure. Casualties are possible	Major damage to infrastructure is likely. Casualties are possible
Advice		Ensure you access the latest risk forecast	Remain vigilant and ensure you access the latest risk forecast	Remain extra vigilant and access the latest risk forecast. Be aware of risks that might be unavoidable. Follow any advice given by	Remain extra vigilant and access the latest risk forecast. Follow orders and any advice given by authorities under all circumstances

				authorities	and be prepared for extraordinary measures

The entire classification procedure that could give rise to a scheme such as shown in **Table 1** relies on cooperation between the end-users and the modellers in determining the forecast accuracy, reliability and specific actions that are recommended in each risk regime. By classifying the risk forecast into different regimes, it is then possible to identify and plan for specific reactions. This ability to prepare is crucial when attempting to deploy probabilistic forecasts in practice.

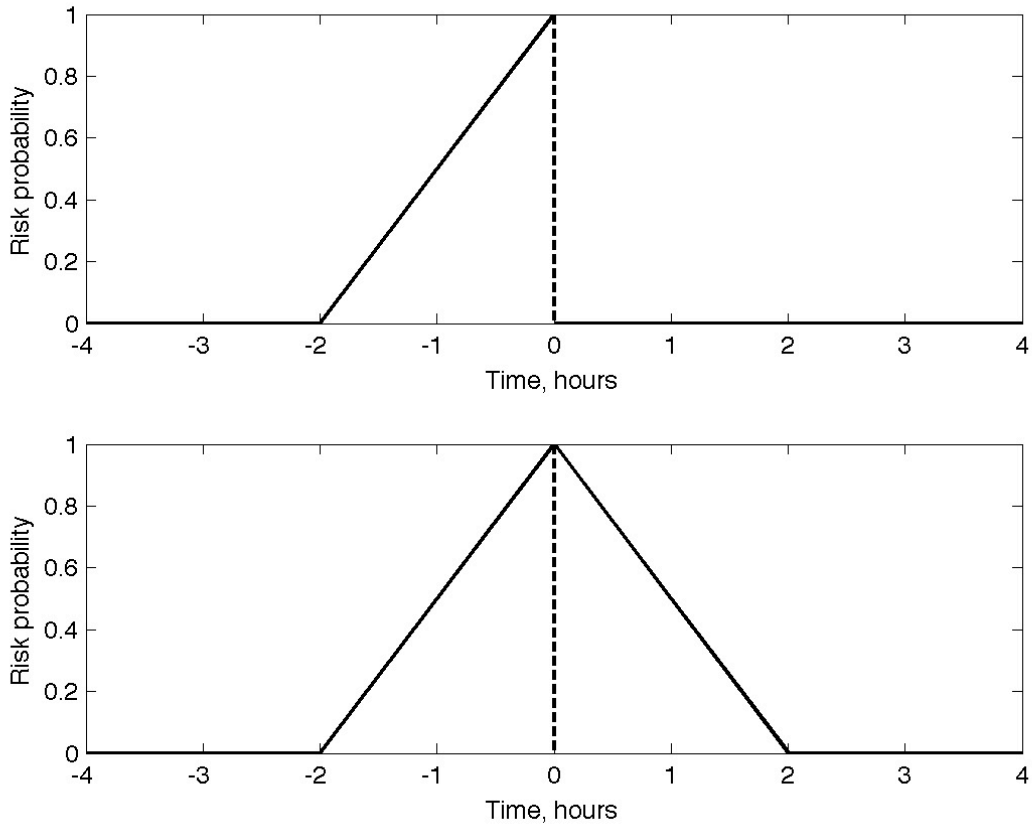


Figure 4: Risk probability labelling before and after an extreme event that occurred at time zero for training a classifier to forecast (top) and detect (bottom) events.

It is also necessary to decide how to measure the accuracy of the early warning system in the epoch immediately after an extreme event. Given that an event has just taken place, the accuracy of the early warning system immediately following the event could be argued to be irrelevant in practice. Given the fact that any feature that is capable of identifying the extreme event requires some time to return to a lower risk level, one possible choice is to employ a linear decrease in the required risk level over a specific time period that follows the extreme event (shown as two hours for illustration purposes in **Figure 4**). The labelling appropriate for training a classifier to forecast and detect extreme events is shown in **Figure 4**). Depending on the particular application, such symmetry in the required risk level (the target for the early warning system) may or may not be appropriate when measuring the overall performance of each approach.

3.3 Mapping features to risk probabilities

Suppose that a particular feature is available for determining the likely risk level associated with specific extreme events. Each feature value, x_t, constructed at time t may be employed to provide an estimate of the risk probability, $\hat{p}_{t+1}$, for the next time instant (constrained to lie in the range [0 1]) by the following simple single parameter monotonic transformation:

$$\hat{p}_{t+1} = \left(\frac{x_t}{x_{\max} - x_{\min}} \right)^q .$$

The effect of this simple relationship is illustrated in **Figure 5**.

An alternative approach is to employ a two-parameter transformation known as the Hill function that displays a sigmoidal response. The Hill function is given by

$$\hat{p}_{t+1} = \frac{x_t^n}{x_t^n + \theta^n} .$$

The value of θ controls the location of the step such that $\hat{p}_{t+1} = 0.5$ when $x_t = \theta$ and the gradient of the step increases with increasing n. **Figure 6** shows how this transformation produces different risk levels for a range of values of θ and n.

We found that the performance of the individual features in providing classification and forecasts of the risk levels was improved by employing the Hill function. It was found to be extremely important to be able to control the location at which the step change takes place.

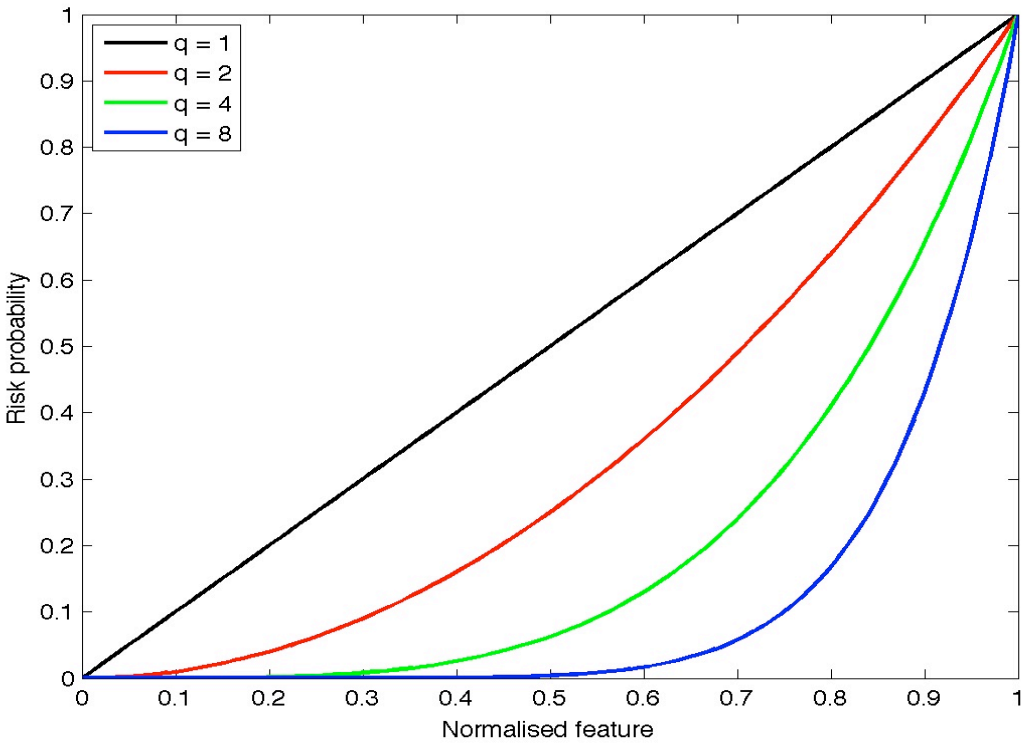


Figure 5: Relationship between the normalised feature and the inferred risk probability depending on the transformation for values of q ranging between 1 and 8. The number of high probability risk alerts decreases as q increases.

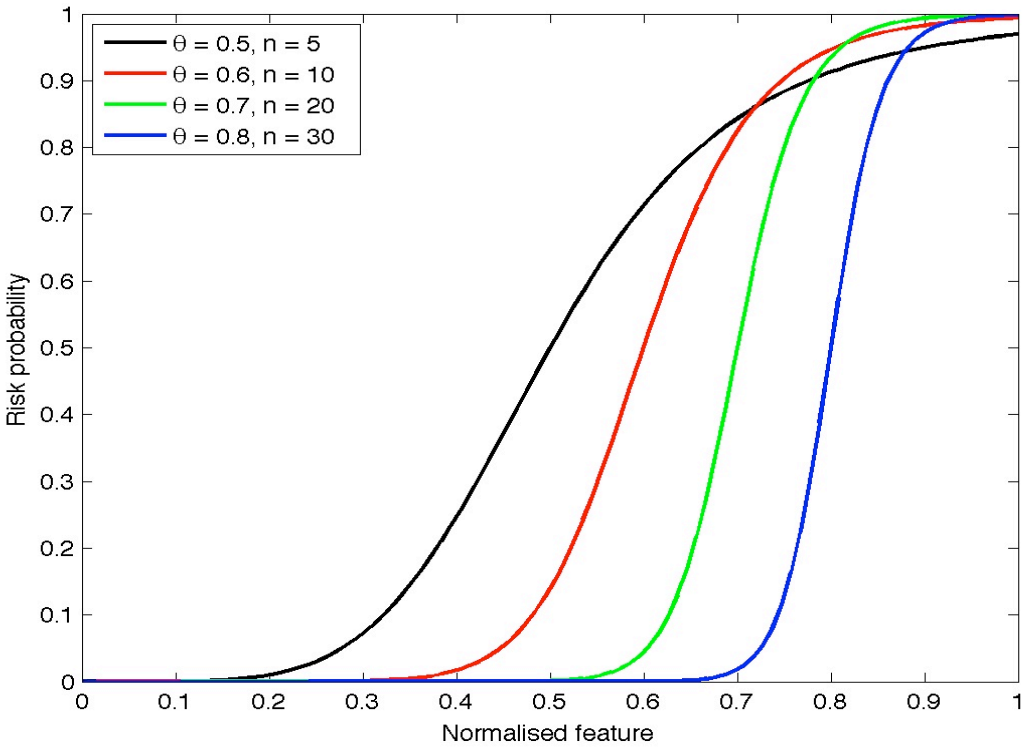


Figure 6: Relationship between the normalised feature and the inferred risk probability depending on the Hill transformation for different values of θ and n.

Classification techniques can be evaluated on the basis of their ability to provide a high level of classification performance on out-of-sample data that was not employed for training the classification techniques. The ability of a technique to deliver accurate classifications can be determined from an ROC analysis of the results where the fraction of true positives, false positives, true negatives and false negatives are reported. These values may be used to graph an ROC curve for each technique showing how the threshold between low and high risk affects the number of false alarms. The area under the curve is often employed as a single measure, from which it is possible to rank and compare the performance of a number of competing techniques.

A classifier may be evaluated using a contingency table, also known as a confusion matrix. Given that an extreme event was either observed or not observed, the corresponding classification is binary. There are four possible outcomes for a binary classifier and these are represented in a contingency table (as shown in Table 2).

Table 2: A contingency table for comparing classifications and verifications

DATA	Observed Extreme	Observed Non-extreme	Total
Classified extreme	h	f	Classified as Extreme Events
Classified non-extreme	m	z	Classified as Non-extreme Events
Total	Observed Extreme Events	Observed Non-extreme Events	N (length of time series)

The four outcomes are as follows: h denotes the number of “hits” (correct extreme alarms); f “false alarms” (incorrect extreme alarms), m “misses” (incorrect non-extreme alarms); and z “zeros” (correct

non-extreme alarms). Note that for a total of N classifications, the entries in the contingency table must satisfy the constraint, N = h+f+m+z.

The probability of detection (POD), also known as hit rate, is defined as,

$$POD = h/(h+m)$$

where h is the number of extreme events (in the augmented extreme event vector) that were classified correctly, while m is the number of number of extreme events that were classified wrongly as non-extreme events. Hence, POD gives an estimate of the fraction of observed events that were correctly classified. POD lies in range [0, 1], with 1 being the perfect score (h=1, m=0). The disadvantage of POD is that it ignores the false alarms and can therefore be artificially improved by issuing more number of classifications as being extreme, so to increase the number of hits (over-classification).

The probability of false detection (POFD), also known as false alarm rate, is defined as,

$$POFD = f/(z+f)$$

POFD gives the fraction of false alarms. It lies in range [0, 1], with 0 being the perfect score (f=0, z=1). The disadvantage of POFD is that it can be artificially improved by issuing fewer extreme event classifications to reduce the number of false alarms.

The receiver operator characteristics (ROC) analysis provides both a visual and quantitative assessment of the performance of a classification technique. The ROC diagram shows the variation in the value of POD versus POFD achieved when changing the parameter(s) of the classifier. The particular application of the classifier is typically used to decide on the most appropriate settings for operating the classifier. This will require finding a suitable balance between increasing POD while decreasing POFD.

When assessing the performance of a binary classifier, a popular choice is the true skill score (TSS). The advantage of using TSS is that it captures the classifier performance by measuring both the probability of detection (POD) and the probability of false detection (POFD), and is given by,

$$TSS = POD – POFD.$$

Taking both POD and POFD in account to compute a final performance score overcomes the individual limitations of both POD and POFD.

4. Spatial distributed forecast uncertainty information

UNIOL has developed an approach to communicate forecast uncertainty for regions, e.g. control zones that are characterized by an inhomogeneous distribution of wind power capacities. Much work has been done in the past to estimate forecast uncertainty and prediction risk intervals for either single sites (e.g. offshore wind parks (Pinson et al., 2009)) or dynamic prediction intervals on the national level (Dobschinski et al., 2010). But even if the aggregated forecast uncertainty is known, end-users (e.g. the transmission system operator) might be interested how the forecast uncertainty is distributed geographically. This knowledge can be used to discriminate regions with high and low uncertainty. Thus, for example, it will be possible in the future to allocate preferably reserve power in areas that exhibit a high risk that wind power forecast are imperfect. Or, in case wind power production and demand are displaced geographically, reserve power can be obtained closer to the sites of demand. In future power systems with Smart Grid technology, i.e. distributed storage, PV and biogas systems and demand side management the uncertainty in wind power production will be important information to develop economic balancing strategies.

The ensemble wind power forecasts used in Task 6.6 base on UNIOL's work in Workpackage 5. The details on the wind power forecast model are given in Deliverable Dp-5.10 (von Bremen et al., 2012). The ensemble wind power forecast utilize the Ensemble Prediction System (EPS) of ECMWF and the

new 100m winds that became operational on 26 January 2010. The results presented here are for the 50Hertz control zone in Eastern Germany. It is worthwhile to mention that real wind power production data with a spatial resolution similar to the EPS (0.25°) does not exist. Thus, 100m winds from the ECMWF analysis are utilized to compute the spatially resolved wind power production. The configuration of the wind power forecasting model is identical to the forecasting mode.

As an indicator of forecast uncertainty the RMS spread, i.e. standard deviation of the wind power ensemble is used. Testing has shown that the 70% inner quantile range of the ensemble distribution is for higher lead times too sensitive with respect to the average forecast error. Figure 7 shows the RMSE of the ensemble mean wind power forecast (solid line) with forecast lead time. The average ensemble spread of wind power (dashed line) is slightly smaller which indicates too little spread (Leutbecher and Palmer, 2008). However, the two lines have the same slope and do not diverge. In contrast to this, the 70% inner quantile range of the ensemble distribution increases steeper with lead time compared to the RMSE of the forecast. Conclusively, the 70% inner quantile can be regarded to be too sensitive for higher lead times.

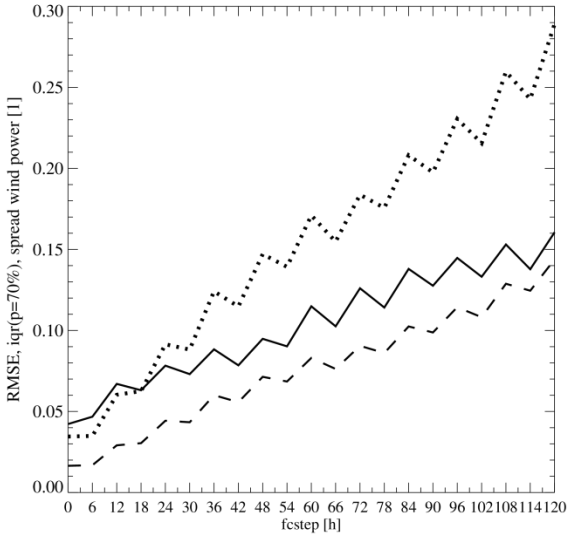


Figure 7: RMSE of ensemble mean (solid), ensemble spread (dashed) and inner quantile (p=70%) of the distribution of ensemble wind power forecasts for 50Hertz in normalized units (to installed wind power) for February and March 2010.

For the occasion of winter storm Xynthia that hit Germany 29. February and 1 March 2010 the spatial distribution of wind power uncertainty in the 50Hertz control zone is shown and discussed here. The powergrams in Figure 8 show that there are clear indications of Xynthia 4 days in advance for the 50Hertz control zone. The 48h wind power forecast of Xynthia (Figure 8, right) is very good in every sense, i.e. the amplitude and the timing is correct for the deterministic and the ensemble forecast. Furthermore the uncertainty is very low as indicated by the short bars.

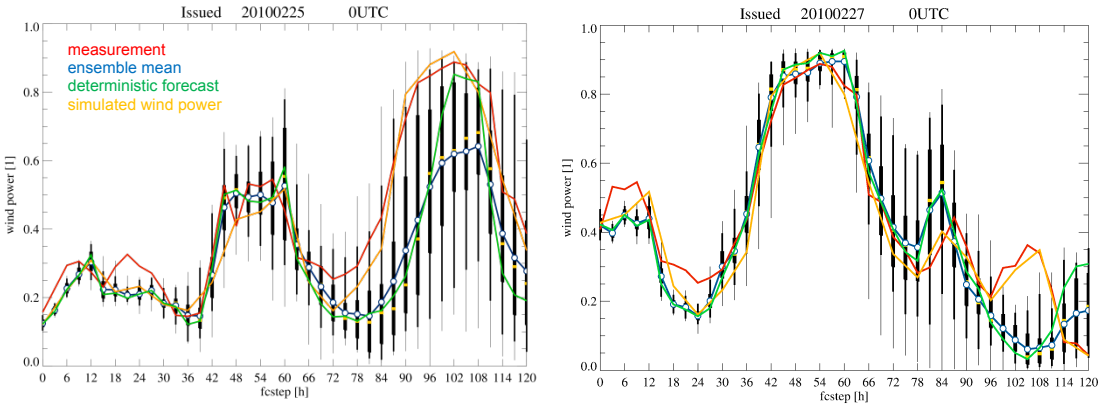
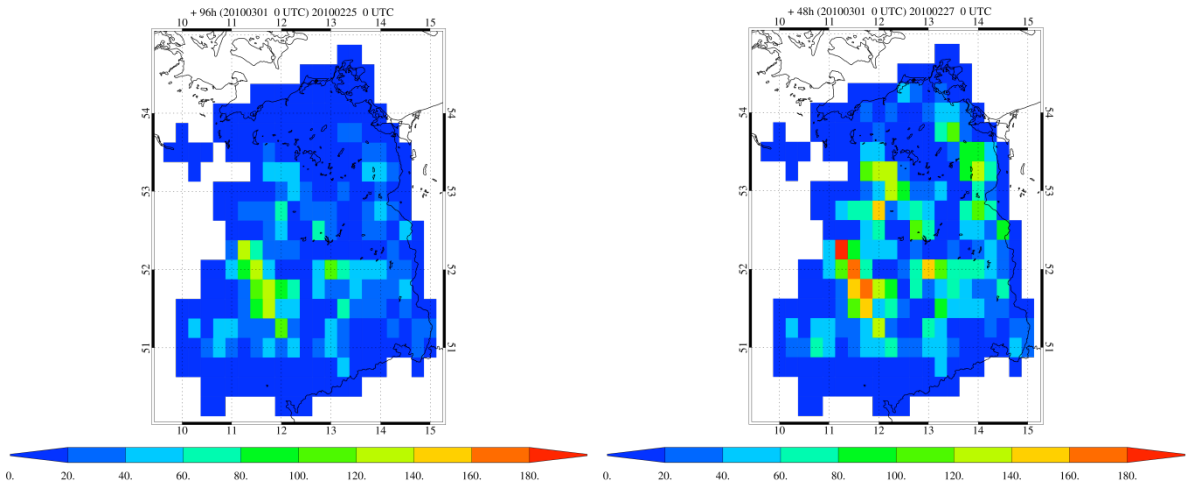


Figure 8: Powergram of probabilistic wind power forecast of storm Xynthia at 1 March 2010 (0UTC) in the 50Hertz control zone. The forecast uncertainty is extremely reduced in the newer forecast run (right) issued at 27 February (0UTC) compared to the medium-term forecast (left). Measured wind power in red, ensemble mean in blue, control (deterministic) forecast in green and simulated wind power in orange. The vertical boxes represent the 50% and 90% inner quantiles while the minimal and maximal value of the ensemble is indicated by the tip of the vertical thin line.

The spatial distribution of the poor 96h wind power forecast of Xynthia can be seen in Figure 9 (left). Large absolute forecast errors of more than 90 MW per grid point occur (Figure 9, bottom left). However, the ensemble spread (Figure 9, middle left) in those grid points is very large. Conclusively, a high uncertainty was indicated and enough time was given to take actions like for instance to increase the amount of regulating reserves to balance deviations from the day-ahead forecast. But as seen in Figure 9 (right bottom) the 48h forecast is far better and the maximal forecast error is less than 30MW per model point. This good forecast skill is anticipated from the low forecast uncertainty, i.e. ensemble spread in Figure 9 (middle right).



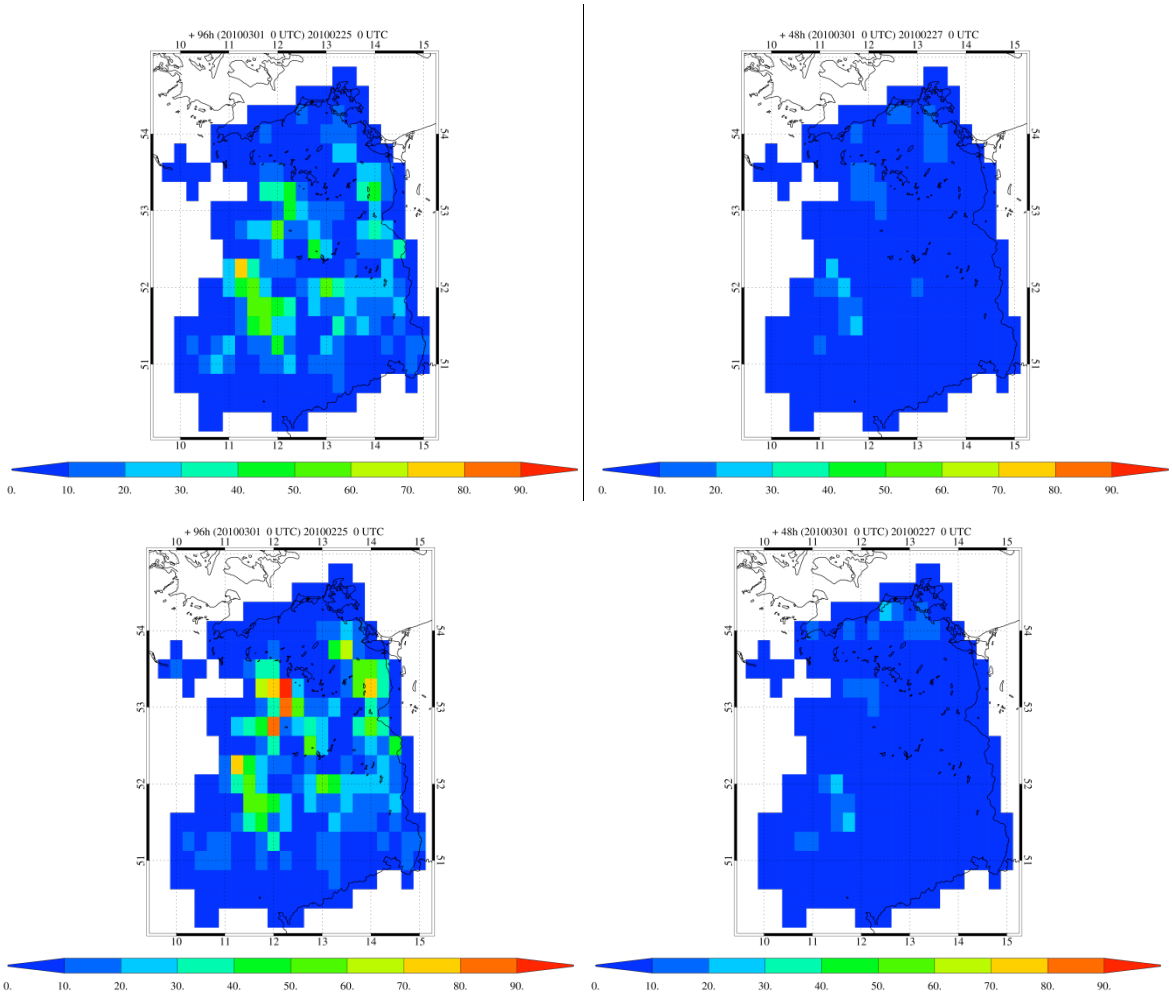


Figure 9: 96h (left) and 48h (right) wind power forecast of storm Xynthia for the 50Hertz control zone valid for 01 March 2010, 0UTC. Ensemble mean (top), forecast uncertainty expressed as ensemble spread (middle) and absolute forecast error (bottom) in MegaWatt.

The scatter plot between the absolute forecast error and the ensemble spread at the level of model points in the 50Hertz control zone is shown in Figure 10 for the +96h (plus signs) and the +48h (bullets) forecast horizon. For both forecast horizons the relation between forecast error and EPS spread is very similar. The occurring values of the EPS spread are much smaller for the +48h forecast and thus it is anticipated that the forecast error is considerably smaller.

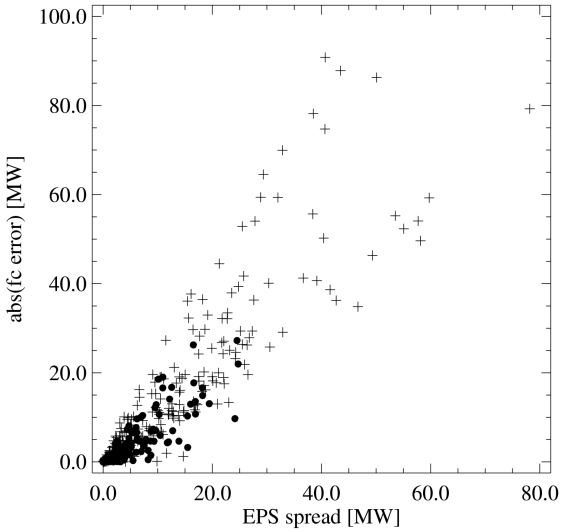


Figure 10: Relation between forecast error and ensemble spread for each model point in the 50Hertz control zone for a +48h (bullets) and +96h (crosses) forecast valid at 1 March 2010 0UTC.

5. Summary

The challenge of communicating the output of complex models to decision-makers and policy-makers is not a new one. Arising from the sophisticated physical and statistical models being developed as part of SafeWind is a need to ensure that the resulting probabilistic forecasts can be conveyed to a diverse group of end-users. While addressing this challenge requires ongoing interaction between modellers and end-users, in this project we have identified some key contributions to facilitate the process.

Probabilistic forecasts in a raw format are likely to be particularly daunting for most non-statisticians. Rather than attempt to convey a full curve, which reflects the forecast density, it may be better to focus on assessing the probability of particular extreme events leading to a binary output. An alternative approach is to present the probabilities through a simple risk classification scheme, which leads to four colour-coded risk levels (green, yellow, amber and red). The advantage of the latter is that it preserves the quantitative information supporting the risk assessment and also allows the end-user to map these risk levels onto previously decided actions. It is this ability to prepare and plan for extreme events which is generally of great interest to decision-makers.

A new approach to communicate ensemble predictions to end-users has been developed. The uncertainty in wind power forecasts is given on a geographic map and enables end-user to spot areas with high forecast uncertainties immediately. The approach is demonstrated for the 50Hertz control zone and as an example the winter storm Xynthia is shown. The RMS ensemble spread has been utilized to represent the uncertainty and is linked very closely to the observed (here simulated) wind power forecast error on NWP model level. While the +96h ECMWF EPS forecast of Xynthia was still quite poor, the +48h forecast of Xynthia is very good.

6. References

G. Anastasiades and P.E. McSharry (2012). Quantile forecasting of wind power using variability indices. (in preparation).

L. von Bremen et al. (2012). Impact of various ensemble configurations for wind power predictions. Deliverable Dp-5.10 in EU-Project “SafeWind” (under preparation)

Davy, R., Milton, J., Russell, C., & Coppin, P. (2010). Statistical downscaling of wind variability from meteorological fields, Boundary-Layer Meteorol, 135, 161–175.

P. Pinson, H.A. Nielsen, H. Madsen, G. Kariniotakis (2009). Skill forecasting from ensemble predictions of wind power. Applied Energy **86(7-8)**, pp. 1326-1334.

P. Pinson, P.E. McSharry, and H. Madsen (2010). Reliability diagrams for density forecasts of continuous variables: accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society* 136(646): 77-90.

J. Dobschinski, Dr. B. Lange, Dr. K. Rohrig, Y. Saint-Drenan, A. Wessel, L. von Bremen (2009). The potential of advanced shortest-term forecasts and dynamic prediction intervals for reducing the wind power induced reserve requirements. In Proc. of EWEC 2010, Warsaw.

A. Lau and P. E. McSharry (2010). Approaches for multi-step density forecasts with application to aggregated wind power. *Annals of Applied Statistics* 4(3): 1311-1341.

P.E. McSharry, P. Pinson and R. Girard (2009). Methodology for the evaluation of probabilistic forecasts. Deliverable Dp-6.2 in EU-Project “SafeWind”.

P.E. McSharry (2010). Should we switch to a new way of predicting weather? BBC Newsnight, 7 January 2010. http://news.bbc.co.uk/1/hi/programmes/newsnight/8448617.stm

P.E. McSharry (2010). Revealing the real forecast. The Guardian. 19 January 2010.

Leutbecher, M., and T. Palmer (2008), Ensemble forecasting, Journal of Computational Physics, 227(7), 3515–3539.

Part 2

Contents

Chapter 1

Introduction

This introduction gives a background to Wind Power Forecasting (WPF) where the needs for forecasting and the usefulness of forecast ensembles are briefly described. This is followed by an outline of the purpose, objective, scope and structure of the report.

1.1 Background to wind power forecasting

As a result of the increased concern about the drawbacks of conventional energy sources in terms of emissions of Carbon Dioxide (CO_2) and other Green-House Gases (GHG), renewable energy sources such as wind power attract more and more attention from governments, industry, research and the general public. Wind power, which after manufacture and installation of the turbines does not emit any GHG, has developed from a small emerging technology some decades ago to an important contributor to the electricity supply. In Denmark for example, more than 20 % of the electricity consumed today is produced by wind power. In Europe, the total installed capacity has increased by a factor of 100 since the beginning of the 1990's [9] and wind power was in 2008 the largest growing electricity source beating coal, hydro, nuclear and photo voltaic. Worldwide, the total installed capacity has exhibited an annual growth of almost 25 % during the last years, and it passed 120 GW in 2008 [9, 15]. Slightly more than half of this capacity is installed in Europe. Denmark, Spain and Germany have been the leaders but other countries such as France, Italy and the UK exhibit an extensive growth. Wind turbines have until know mostly been installed onshore but offshore development is growing fast [10].

This tremendous increase in installed capacity has partly been made possible by subsidies from governmental bodies in many countries. The development is expected to continue in the coming decades, encouraged by the European Commission binding target of 20 % of energy to be covered by renewable sources by 2020 [9]. The wind power sector has, like many other industries, been affected by the financial crisis. The impact is however lower and there is strong optimism in wind power as an important part of the solution towards a more sustainable energy system and a sustainable future [10].

1.1.1 The needs and benefits of forecasts

Wind energy is an inexhaustible but variable energy source. The variability is due to the seasonal, daily and regional variations in wind speed. The power output from wind turbines is therefore variable and managing its integration into the power system is a rather complicated issue for Transmission System Operators (TSOs). This becomes increasingly challenging when larger parts of the power supply are covered by wind power. In order to integrate the power efficiently into the grid, accurate forecasts of the expected wind power output for the coming hours and days are necessary.

WPF therefore constitutes an important tool in managing the integration of wind power production. Different electricity sources, such as coal plants, hydro stations and so forth,

have different ramp rates and thus different possibilities to adjust to forecast errors from wind power. Increased forecast accuracy reduces the costs for owners of wind farms and facilitates the integration of wind power into the electricity grid. For electricity systems that rely to a large extent on carbon intensive fuels such as coal, a better integration of wind power into the grid also contributes to reduced CO_2 emissions. The economic benefits of using wind power forecasts vary depending on how the forecasts are used and which share wind power has in the power system. Studies show that savings of several € cents per kWh can be made [13].

Meanwhile, the deregulation of the electricity market in several countries sets additional demands within electricity trading. Accurate wind power forecasts are important for energy traders in order to minimize the costs that deviations between predicted and observed power can imply.

1.1.2 Ensemble forecasts and risk indices

Current WPF models provide end-users with point predictions as well as prediction intervals for a coming time interval. The time intervals considered can vary from a few minutes to several weeks ahead depending on the purpose of the forecast. To ameliorate the integration into the electricity grid, a forecast period of 2 - 3 days ahead is usually considered. Wind power forecasts for this forecast length are based on historic power measurements and Numerical Weather Predictions (NWPs) provided by meteorological services.

Due to the chaotic behaviour of the atmosphere, NWPs are associated with errors which can grow in time and space. These errors are amplified or dampened when the relevant variables, mainly wind speed and direction, are used in WPF models. Nowadays, weather forecasts are often provided as meteorological ensembles derived from, for example, perturbations of the initial conditions in a NWP. The ensembles contain a number of alternative scenarios and the dispersion among those is expected to reflect the uncertainty of the forecast. A measure of this dispersion, based on the standard deviation of the ensemble members, was introduced in [22]. It was named Normalized Prediction Risk Index (NPRI) since one of the purposes of its use was to give information about the expected risk for large errors in a forecast. Information about the uncertainty of the forecast and the risk for large errors can serve as a complement to point predictions and constitute very valuable information for decision makers.

1.2 Purpose and objectives of the work

The concept of risk indices for wind power forecasts was introduced as a part of the European Commission project ANEMOS [2] and is under investigation in other research activities performed in the area. This report work has been conducted in the frame of the two ongoing European Commission projects SafeWind [4] and ANEMOS.plus [3] which are coordinated by the Centre for Energy and Processes (CEP).

Using the ideas developed in [22] as a starting point, the purposes of this work are to further investigate the usefulness of risk indices derived from wind power forecast ensembles. The analysis is partitioned into a number of different tasks which develop the previous findings in different directions.

- In [22], the NPRI was evaluated on a Danish offshore wind farm. In order to investigate whether similar results are obtained in other cases, the first purpose of this work is to apply this definition on three French onshore wind farms.

- The NPRI was evaluated on a time window of 24 hours, either for day 2 or day 3 ahead. The second purpose of this work is therefore to evaluate the performance of the index for other lengths and locations of the time window.

- Since the focus so far has only been on a single wind farm, the usefulness of the NPRI when forecasting for several wind farms is also investigated.

- There is however nothing that confirms that the existing definition of NPRI is the most suitable method to extract information from ensemble forecasts in order to provide information about the uncertainty and the risk for large forecast errors. Modifications of the NPRI and a number of alternative risk index measures are therefore considered.

- The way risk indices are presented to end-users is also of concern and alternatives to the previously used presentation technique are investigated to check whether more suitable options exist.

In order to compute risk indices from wind power ensembles, it is however first necessary to produce such ensembles. This can be done in several ways using more or less advanced WPF models. One of the main challenges is how the WPF models preserve the ensemble spread when the relevant NWP ensemble variables are converted to power ensembles. Two WPF models, one basic and one more advanced, are therefore used to derive wind power ensembles from NWP ensembles in order to compare their performance.

Before carrying out this, detailed introductions to WPF, ensemble forecasting and risk indices are made with the purpose of constituting a solid basis for the remainder of the work.

1.3 Scope

The project includes modelling of wind speed to wind power using point prediction WPF models. Probabilistic forecasts, which give point predictions together with prediction intervals, are not considered.

There exists numerous advanced WPF models developed by research centres and industry actors but only one advanced model is used in this work. The reason for this is that the scope of the work is not to produce as accurate forecasts as possible but rather to extract as much information as possible from the dispersion of the ensembles. The associated possible improvements which can be obtained in WPF models are therefore not examined. Even though the scope is neither to investigate whether the mean of the power ensemble members can be a more accurate forecast than any of the ensemble members, this is briefly examined in order to compared with previous results.

1.4 Structure of the report

Chapter 2, presents the WPF models employed in this work to generate wind power forecast ensembles. The wind farms and data used in the work are also described and the performance of the models is then evaluated. The results of the model performance are discussed followed by a presentation of some of the characteristics found in the ensembles and the models.

With this as a base, risk indices can be evaluated on the generated wind power forecast ensembles. This investigation is made following the purposes presented previously in Section 1.2. It starts with an investigation of the NPRI on the three French wind farms in Chapter 3 including a comparison with the results obtained in [22]. This chapter also contains an extension of NPRI to other temporal and spatial scales.

In the subsequent chapter, Chapter 4, a number of new concepts are introduced and this chapter therefore contains the main contributions of this work in the field of WPF. Two modifications of the NPRI and a number of alternative risk indices are defined and evaluated. This chapter also includes an investigation and discussion of how risk indices can be presented.

Finally, the report ends with conclusions and perspectives for future work in Chapter 5. This is followed by three appendices, Appendix A, B and C where supplementary results are presented.

1.5 Ensemble forecasting

The accuracy of the NWP data is critical when producing wind power forecasts but this is not in the hands of wind power forecasters. Meteorological ensembles representing different future situations can therefore be an interesting alternative to use within WPF, for example as a mean to obtain wind power forecast ensembles.

1.5.1 Meteorological ensembles

Variations in the initial conditions or the formulation of NWP models can cause large deviations in the output. Besides increasing the spatial and temporal resolution, effort from meteorological institutions is therefore also placed on providing *meteorological ensembles* which consist of sets of alternative forecasts representing different scenarios. Information about possible alternative outcomes can be of importance for many different NWP users, especially to determine alternative strategies depending on the probability for different weather scenarios.

ECMWF uses its Ensemble Prediction System (EPS) to provide meteorological ensembles consisting of 51 ensemble members. Out of these members, one member is an unperturbed *control forecast* and the rest are alternative predictions. The ensembles are calculated using *singular vectors* which are the most perturbed states in terms of energy growth in the model during the first two days ahead [8]. The ensemble members are then calculated as linear combinations of the vectors, which imply that the largest possible deviations are taken into account. The model also integrates a stochastic representation of model uncertainties.

NCEP uses a different method, called the *breeding method*, to produce 11 ensemble members. The method consists of adding so called bred modes to the control forecast. These are obtained by cyclically perturbing and rescaling the control forecast and imply that the perturbations grow along the forecast trajectory [8].

Other ways of constructing meteorological ensembles are to use different meteorological models or different parameterizations of a model but with the same initial conditions.

1.5.2 Wind power forecasts ensembles based on NWP ensembles

A simple way to produce wind power ensembles is to use some of the meteorological ensembles presented in the previous section and feed them into a WPF model to obtain a set of point or probabilistic predictions. Using a statistical WPF model, the model can be trained on the control member and forecasts can thereafter be made for all ensemble members. The mean and the standard deviation of the ensemble members can also serve as explanatory variables in WPF models. Another approach to obtain a set of different wind power predictions is to use different WPF models with the same meteorological data [14].

Ensemble forecasts, either of wind speed or of wind power, can be visualized using so called *spaghetti plots* which show the different ensemble members for a look-ahead time window. An example is shown in Figure 1.1 which shows point wind power forecasts derived from the 51 ensemble members from an ECMWF meteorological forecast. The control member and the mean of the ensemble members are marked specifically and the observations are also included.

In the episode presented here, the ensemble members lie relatively close for the first 30 - 40 hours after which the spread is larger for the rest of the look-ahead time window. An interesting observation is that the peer order of the ensemble members does not change much for some intervals, for example between 25 and 40 hours, while it changes significantly in other time intervals.

Once wind power forecast ensembles are obtained, statistics such as the mean and standard deviation of the ensemble members can be derived. Deriving point predictions by calculating the mean of the power ensemble members for each look-ahead time has been found to reduce the average forecast error for both ECMWF and NCEP ensembles, [22]. The improvement increased with look-ahead time, from a couple of percent for 1 - 2 days ahead up to 10 - 15 % for 4 - 5 days ahead.

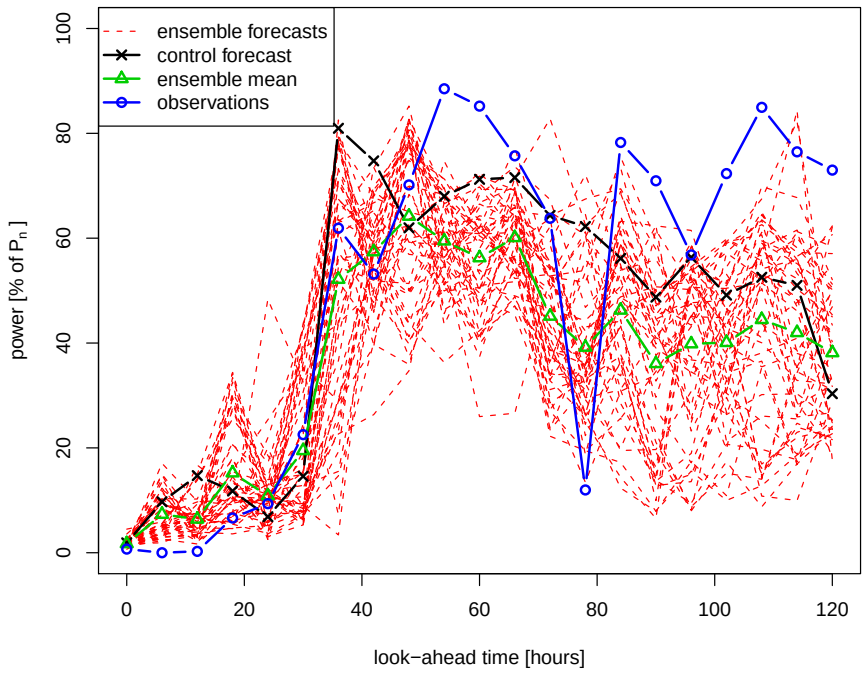


Figure 1.1: Example of power forecast ensembles based on ECMWF meteorological ensembles with associated observations.

1.5.3 Temporal ensembles

Another way to produce wind power forecast ensembles is to use *temporal ensembles*, often referred to as *poor man's ensembles*. The ensemble members in this case consist of point forecasts for a certain period made with the same WPF model and for the same set of look-ahead times but issued at different moments in time. An example of such ensembles is shown in Figure 1.2 where five temporal ensembles, issued with a time step of 24 hours, are shown together with observed values.

In this case, the agreement among the ensemble members is rather high for the first 30 hours ahead, indicating no major changes in the weather during the period. A case where the disagreement among the temporal ensemble members is higher, as from 40 hours and onwards in the figure above, is assumed to be related to higher uncertainty and thus translate to higher risk for large forecast errors.

Temporal ensembles are easy to obtain by simply combining a number of successive point forecasts. They can therefore be a cheaper alternative in order to extract more information from ordinary point predictions. It was found in [22] that a weighted mean computed from temporal ensembles gave a more accurate prediction compared with the last issued forecast. The spread of the ensemble members was also useful in assessing the uncertainty of the situation.

1.5.4 Previous results from the use of ensemble forecasts

It can clearly be seen in the spaghetti plot in Figure 1.1 that the forecast errors tend to be larger when the spread among the ensemble members is larger. This was also found in [22], where it was concluded that the standard deviation of prediction errors increases as the ensemble spread increases. It has however also been found that the spread of the WPF ensemble members does not reflect the whole range of possible outcomes [20]. This trend can also be observed in

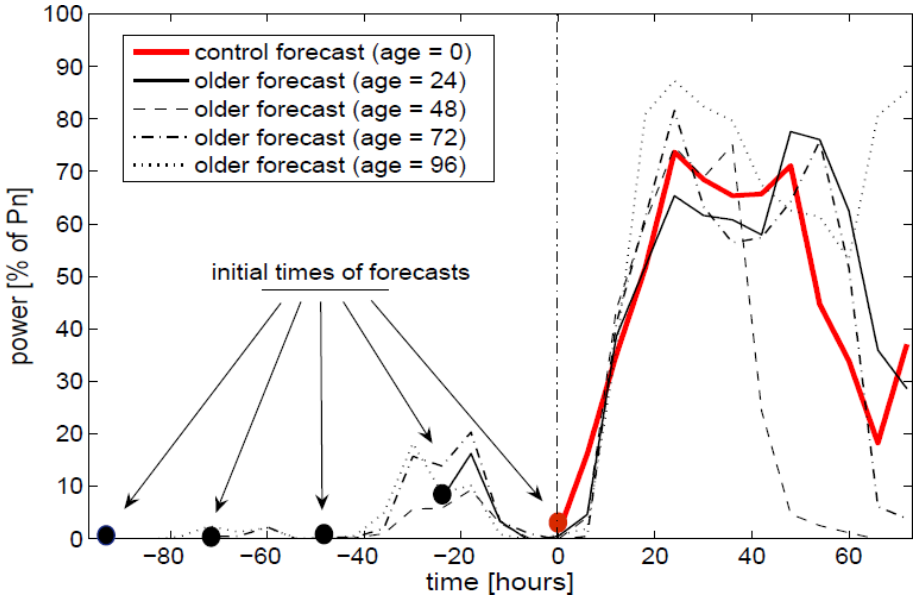


Figure 1.2: Example of temporal ensembles with associated observations. The forecasts are issued with 24 hours in between. The Figure is taken from [22].

Figure 1.1 where the observations are in many cases larger or smaller than the set of ensemble members. This has been explained to be due to a misrepresentation of the uncertainty in the initial state of the atmosphere or to error growth in the model. It leads to underestimation of extreme events, in terms of both size and probability.

A manner to investigate if the ensemble members are representative for possible outcomes is to study whether the observations fall with the same rate over the whole range of ensemble members. This can be done using *Talagrand diagrams* [8] by counting the frequencies of where in the range of ensemble members the observations are found. In order to perform this, the n ensemble members are sorted according to their value for each run and each look-ahead time and the position of the observation among the ensemble members is determined. For n ensemble members, this gives $n + 1$ possible positions for the observation. If the ensemble members where correct in a probabilistic sense, the observations would be found as many times in each of the $n + 1$ positions. For weather prediction ensembles, the diagrams are however generally U-shaped with more observations lying outside the range of the ensemble members [8].

This underestimation of low and high values implies that ensemble forecasts of wind power are not reliable in a probabilistic sense, where reliability is referred to as the probabilistic correctness of the predicted distributions. To provide probabilistic forecasts from ensembles, some kind of recalibration is therefore needed [23]. One method that has been employed is Adaptive Kernel dressing. In [23] a predictive density of wind power was defined as the weighted sum of Gaussian kernels, parameterized by a mean-variance model, associated to each ensemble member. It was shown that the ensembles where more reliable after the recalibration. Another recalibration method was proposed in [14] and more about probabilistic forecasts obtained from ensembles of NWPs is found in [5].

In order to produce wind power ensemble forecasts with as good probabilistic correctness of the members as possible, it is needed to use WPF models that preserve the ensemble spread well [24]. For WPF models that are based on fitting a power curve, this fitting can be made using two different criteria as displayed in Figure **??**. LS fitting of a power curve should be used if the aim is minimizing the RMSE criterion [24]. On the other hand, if the objective is to obtain the most accurate description of the power curve in terms of low bias, orthogonal fitting should be employed. Orthogonal fitting gives a less compressed power curve, thus giving a

power curve that better spans low and high power outputs. The model used in [22] to generate wind power ensembles from ensembles of wind speed and direction uses this idea. This result in larger forecast errors compared to state of the art models but the spread of the ensemble members is better preserved which implies that more information can be extracted when using risk indices.

1.6 Risk Indices

The idea behind risk indices is to capture information from wind power forecast ensembles to give an indication of the expected level of forecast error and the uncertainty of the forecast. They can thus be used to inform on the risk of relying on the forecast. An existing method to do this is the Normalized Prediction Risk Index (NPRI), introduced in [22] and presented here.

1.6.1 Definition of the Normalized Prediction Risk Index

The NPRI was introduced as a measure of the spread of the ensemble members. It is based on the weighted standard deviation $\widetilde{\sigma}_{t,k}$ of the J ensemble members $\widehat{P}^{(j)}_{t+k|t}$ at prediction time t for look-ahead time $t+k$:

$$\widetilde{\sigma}_{t,k} = \left[\frac{J}{J-1} \sum_{j=1}^{J} w_j \left(\widehat{P}^{(j)}_{t+k|t} - \overline{P}^{J}_{t+k|t} \right)^2 \right]^{1/2} \tag{1.1}$$

where the sum of the weights, w_j equals one:

$$\sum_{j=1}^{J} w_j = 1$$

and $\overline{P}^{J}_{t+k|t}$ is the mean of the ensemble members for look-ahead time k:

$$\overline{P}^{J}_{t+k|t} = \frac{1}{J} \sum_{j=1}^{J} \widehat{P}^{(j)}_{t+k|t} \tag{1.2}$$

For ensembles where all scenarios are expected to be equally likely, as in the case of ensembles obtained from ECMWF or NCEP NWPs [23], w_j is set to $1/J$ for all ensemble members. In other cases it can be advantageous to give different weights to different members, for example when concerning temporal ensembles for which the more recent predictions are assumed to better explain the future since they are based on more up to date information.

The NPRI is then defined as:

$$NPRI(k_1, k_2) = \frac{1}{k_2 - k_1 + 1} \sum_{i=k_1}^{k_2} \widetilde{\sigma}_{t,i} \tag{1.3}$$

where k_1 and k_2 are two look-ahead times, $k_1 \leq k_2$.

The NPRI can be given for individual look-ahead times, when $k_1 = k_2$, or look-ahead time windows, $k_2 > k_1$. Following the notations in [22], an NPRI given for individual look-ahead times is noted $NPRI_h$ and an index given over one day, $k_2 - k_1 = 24$ assuming a temporal resolution of one hour, is noted $NPRI_d$. This option gives the average NPRI value of the considered look-ahead time window. The idea behind considering time windows $[k_1, k_2]$ is to take into account the fact that weather and thereby power predictability seldom changes significantly within short time periods due to the relatively slow changes of atmospheric processes.

Using the NPRI to inform on expected energy imbalance

When the NPRI is given as an average over a look-ahead time window, for example as $NPRI_d$, it is also necessary to consider some sort of average of the error over these look-ahead times. In an operational context such as for TSOs, the total energy imbalance D of a certain look-ahead time window, $[k_1, k_2]$, can be of interest. This imbalance is given by:

$$D_{t+k_1}^{t+k_2} = t_r \sum_{i=k_1}^{k_2} |P_{t+i} - \widehat{P}_{t+i|t}| \tag{1.4}$$

where t_r is the temporal resolution of the wind power predictions. The energy imbalance over a look-ahead time window can thereby be given in units of Wh.

Furthermore, it can be advantageous to normalize the energy imbalances by their climatological value for a look-ahead time window. This value is given by the average energy imbalance over a longer time period, for example a year, covering N imbalances:

$$\overline{D}_{k_1}^{k_2} = \frac{1}{N} \sum_{t=1}^{N} D_{t+k_1}^{t+k_2} \tag{1.5}$$

Normalizing the energy imbalances by the climatological value enables a comparison between energy imbalances that are smaller or larger than normal.

It is however not the value of a risk index in itself that gives information about the uncertainty of the situation but it is rather its location in the climatological distribution of risk index values [22]. Sorting and dividing risk index values into a number of classes is a way to compare them with associated energy imbalance levels. Statistics of the observed energy imbalances can then be given for each class in order to give an estimation of the expected energy imbalance corresponding to an obtained index value. The quantiles of the energy imbalance distributions for each class can then be used to give information about the probability for different level of imbalances. This can for example be used to give the probability, or risk, for an energy imbalance larger than 150 % of the average imbalance.

Evaluation of the NPRI

In order to evaluate the performance of NPRI, two criteria are considered; the indices ability to differentiate between high and low levels of energy imbalance and the uncertainty about the expected energy imbalance.

In [22], differentiation ability was defined as the ratio between the mean imbalance level in the last and first class. A high ratio indicates that the index can distinguish well between more and less predictable situations.

Information about the sharpness of the distribution of the level of relative imbalance in each class can be given using the quantiles. A quantile measure that was discussed in [21] is the Inter-Quartile Range (IQR) which is the difference between the upper and lower quartiles, $Q_{0.75}$ and $Q_{0.25}$:

$$IQR = Q_{0.75} - Q_{0.25} \tag{1.6}$$

A lower IQR means that the NPRI gives a sharper indication of the expected energy imbalance while a higher value means that the uncertainty of the expected energy imbalance is larger. How the uncertainty varies between the classes can then be given. An increase of the IQR with class number indicates that the uncertainty about the energy imbalances increases with NPRI value.

Previous results using the NPRI

In [21], a study was carried out using ensembles from both ECMWF and NCEP as well as temporal ensembles. The NPRI values were grouped into 5 equally populated classes and the energy imbalances where normalized by their 10-month climatological mean. The study found

a steady, quasi-linear, increase in mean energy imbalance level for increased NPRI for both $NPRI_d$ and $NPRI_h$ showing that increased spread of the ensemble members is associated to larger average prediction errors. It was also found that the IQR of the energy imbalance distribution increased with NPRI value, indicating that the uncertainty and hence the mean error grows with NPRI value thereby also the risk for larger errors increase.

Power ensembles obtained from both ECMWF and NCEP NWP ensembles were found to have a greater ability to distinguish between more and less predictable situations than temporal ensembles. Overall, NWP ensembles obtained using ECMWF ensembles were slightly better than those obtained using NCEP ensembles.

1.6.2 Presentation of risk indices

Presenting the information extracted from risk indices can be made in several ways. Using classes, the class which an obtained risk index value belongs to, can be given as a measure of the uncertainty or risk level, where the uncertainty ranges from 1 to 5 in the case of 5 classes. The uncertainty can however vary from case to case depending on predictability and the interpretation of different values needs therefore to be adapted to each case.

Although a numeric scale can be useful for the presentation of risk index values, it is not always the most efficient way to show the information. An alternative is to use colour scales with, for example, one colour for each class. Great care in the psychological interpretations of colours must be taken in order to chose the colours appropriately so that they send the right messages [17]. It is therefore important to clearly state the index values and the probabilities associated to each colour.

Uncertainty terminology

When presenting forecast uncertainty it is important to be aware of the fact that uncertainty arises from many sources. Apart from the already mentioned sources such as uncertainty in NWP data and in the wind to power conversion process, there is also uncertainty related to the presentation and interpretation of forecasts.

A certain risk index value can therefore be decoded differently depending on the situation, the users' sensitivity for different uncertainty levels and the users' confidence in the forecaster. The scale chosen and the terminology used when presenting the values is therefore of significant importance. Messages of uncertainty levels such as "high probability", "low risk", "possible" and so forth can be interpreted in several ways. Their definition and link to probabilities must therefore be stated in order to avoid misunderstandings [17]. The World Meteorological Organization (WMO) proposes a scale that could be used to communicate forecast uncertainty. The scale, presented in Table 1.1, is based on the *Likelihood scale* developed by the International Panel on Climate Change (IPCC), and defines the most common uncertainty terms [17].

Table 1.1: Forecast Likelihood Scale [17].

Terminology	**Likelihood of the occurence/outcome**
Extremely likely	Greater than 99 % probability
Very likely	90 % to 99 % probability
Likely	70 % to 89 % probability
Probable - more likely than not	55 % to 69 % probability
Equally likely as not	45 % to 54 % probability
Possible - less likely than not	30 % to 44 % probability
Unlikely	10 % to 29 % probability
Very unlikely	1 % to 9 % probability
Extremely unlikely	Less than 1 % probability

This scale could be a starting point when defining the terminology associated with the presentation of uncertainty of wind power forecasts. It could for example be used when informing on the level of risk for energy imbalances larger than usual. The scale could then easily be adapted using the information provided on the distribution of energy imbalances for a given value of NPRI. The choice of the precise numbers and levels must however be carried out with care.

1.7 Conclusions

This chapter has presented the basics behind wind power forecasting and outlined its characteristics and challenges, especially with respect to the characteristics of the wind-power curve. How the power output from wind turbines can be forecasted, based on for example predictions of wind speed and direction, has been explained and how to evaluate the performance of forecasting models has been presented.

This chapter has also described how wind power forecast ensembles can be obtained in different ways. Furthermore, it has been discussed how the dispersion among the ensemble members can be linked to the forecast error and the uncertainty of the prediction. Since different WPF models preserves the ensemble spread differently, as discussed in Section 1.5, it is important for this study to compare the results when using WPF models with different properties.

After this chapter, that has provided a solid base in WPF, ensemble forecasting and risk indices, the principal work can begin. This starts with a detailed description of the WPF models employed in the work in the next chapter including information about the wind farms and the data. The models are also evaluated and some important characteristics found in the models and the ensembles are presented.

Chapter 2

Description and Evaluation of the Wind Power Forecasting Models

Wind power forecast ensembles can be generated in several ways, as discussed in the previous chapter. In this work, four options are considered: an advanced statistical point prediction model, a model based on the theoretical power curves, temporal ensembles and a new approach. The new approach proposed in this work is named *multi ensembles* and consists in combining standard ensembles and temporal ensembles giving a larger set of ensemble members. How the ensemble members are generated in the four cases is described in the beginning of this chapter.

This work is carried out on three French onshore wind farms and this chapter continues with a detailed presentation of the farms. The data used as input in the WPF models is also presented in detail. Furthermore, the performance of the prediction models is evaluated and discussed. The chapter finishes off with a deeper analysis of some of the characteristics found in the ensemble members and the models which are of importance for the remainder of the work.

2.1 Producing wind power ensembles

The meteorological ensembles employed in this work come from ECMWF. The four methods to generate wind power forecast ensembles have different characteristics and use these meteorological ensembles differently.

2.1.1 Using an advanced statistical point prediction model

The advanced model employed in this work to produce wind power predictions is a statistical model called *Random Forest* (RF). RF is a regression method that uses classification trees to establish the relation between the observations and a set of explanatory variables in order to make predictions. The model is tuned to minimize the forecast error in terms of RMSE [19]. More information about the model can be found in [6] and [19].

The wind speed and direction at 10 meter above ground level (a.g.l) are used as explanatory variables together with measurements of the wind power output at look-ahead time 0. The control member of the NWP is used in the RF model to obtain the relationship between these variables and the power output. This relationship is then used to predict the power for the control member. The alternative ensemble predictions are assumed to have the same properties as the control member and the trained model can therefore be used on all NWP ensemble members to produce power ensembles.

When using statistical WPF models such as RF that need model training, it is important to perform training and testing of the model on different and independent data sets. Data containing NWP and previous measurements are therefore split into two sets, with a portion of 2/3 of the available time-series used for the learning part and the rest for the testing. The model is tuned on the learning set without considering the testing set on which its performance is then evaluated. A schematic view of the modelling procedure, including the division into learning and testing sets, is shown in Figure 2.1.

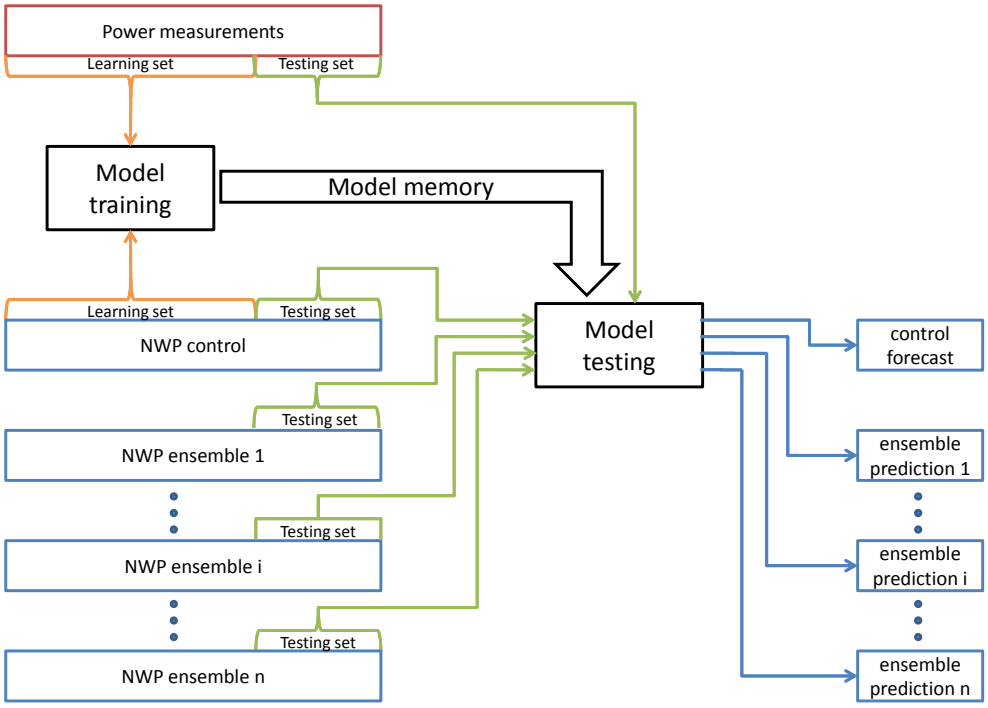


Figure 2.1: The procedure of modelling wind power ensembles using a statistical WPF model.

The model is trained and tested for each look-ahead separately since the relations between the explanatory variables vary as a function of look-ahead time.

Other meteorological variables than the two used here, such as air density, temperature, pressure, and humidity, could also be used as explanatory variables. Several studies show however that the benefits of incorporating these variables are small, with the drawbacks of increased computational cost. Since the scope of this study is not to produce as accurate forecasts as possible, it was decided to only use wind speed and direction.

2.1.2 Using a model based on theoretical power curves

A prediction model with a low bias with respect to the power curve should be employed in order to better preserve the ensemble spread. To compare with the results obtained with RF, predictions are therefore also made with a *Power Curve* (PC) model. This is a rather simple physical model that uses the relationship between wind speed and power in the theoretical power curves for the farms to predict the power output from wind speed predictions. The advantage is that no statistical learning is made and the PC model can therefore be directly applied on all the NWP ensemble members. The model is however expected to produce larger forecast errors since it it is known that a theoretical power curve is not ideal for modelling the power output.

Theoretical power curves of the turbines are provided in [16] for the three farms with a resolution of 1 m/s. They are provided at hub height for each turbine individually and are therefore aggregated to get an approximation of the total power curve for each farm. Since the theoretical power curves are given for wind speeds at hub heights, the predicted wind speeds are first downscaled to hub height, using Equation **??** with values of z_0 and z_2 taken from [16].

2.1.3 Obtaining temporal ensembles

Temporal ensembles are obtained by combining the control members of the power ensembles produced by the RF model or the PC model. The time step between two consecutive NWP forecast runs is 12 hours. To produce temporal ensembles for look-ahead times from 0 to 72 hours, which cover the first three days that are of main interest in this work, 5 forecast runs are combined as shown in Table 2.1.

Table 2.1: Combination of 5 forecast runs with 12 hours between the runs to obtain temporal ensembles for look-ahead times from 0 to 72 hours ahead.

Forecast age	**Look-ahead times used**
0 h	0, 6, 12, ..., 60, 66, 72
12 h	12, 18, 24, ..., 72, 78, 84
24 h	24, 30, 36, ..., 84, 90, 96
36 h	36, 42, 48, ..., 96, 102, 108
48 h	48, 54, 60, ..., 108, 114, 120

2.1.4 Obtaining multi ensembles

Constructing temporal ensembles from standard power ensembles allows increasing the dimension of the ensemble set by combining the two parts into a larger set of ensemble members. This proposal is defined here as *multi ensembles* and is expected to be a way to capture as much information as possible from the predictions. A larger set of ensemble members is also assumed to give a higher probabilistic correctness when analysing the spread.

Using 51 power ensemble members and 5 temporal ensembles per member, this gives 255 multi ensemble members. Multi ensembles can, in the same way as temporal ensembles, be constructed from both the ensemble forecasts made by the RF model and those made by the PC model.

2.1.5 Weighting of temporal and multi ensembles

In the case of power ensembles obtained from ensembles of NWP, all the different scenarios are assumed to have similar properties and all ensemble members are therefore given the same weights. Considering temporal ensembles on the other hand, the most recently issued forecast contains more up to date information and is therefore assumed to be the most accurate one. It is therefore reasonable to give more weight to the more recent predictions when computing the weighted mean or the weighted standard deviation of the ensemble members.

The choice of weights is delicate and needs to be optimized according to some criteria, for example lowest average NRMSE or largest difference in mean energy imbalance between classes of NPRI. In [22], temporal ensembles of ages 0, 24, 48, 72 and 96 hours were used. They were given the weights 0.71, 0.1, 0.07, 0.06 and 0.05 respectively. Those weights were found empirically and a drawback with this is that no consistency is guaranteed. For comparability purposes it is therefore motivated to standardize a way to obtain weights for temporal ensembles. It is reasonable to assume that the accuracy of the forecasts decreases exponentially in

time. Using this proposition, the weights for a forecast aged t can then be calculated as:

$$w_t = \frac{\exp^{-t/\beta}}{\sum_{s=(0,12,24,36,48)} \exp^{-s/\beta}} \tag{2.1}$$

where β is a scaling parameter with unit time and t takes the same values as s. The parameter β can then be varied to give different weight decay. A lower value gives more importance to the latest issued forecasts and a higher value gives more similar contribution to all members. This is shown in Table 2.2 where the weights for β equal to 12, 24 and 48 are shown.

Table 2.2: Example of exponentially decaying weights that can be used when weighting temporal ensembles.

β	**Forecast age = 0 h**	**Forecast age = 12 h**	**Forecast age = 24 h**	**Forecast age = 36 h**	**Forecast age = 48 h**
12	0.636	0.234	0.086	0.032	0.012
24	0.429	0.260	0.158	0.096	0.058
48	0.310	0.241	0.188	0.146	0.114

Using this equation, the weight distribution is determined by a single parameter β. The parameter s depends on the number of ensemble members that it is decided to include, a choice that can have an impact on the results.

The weights used for the temporal ensembles can be used for multi ensembles as well, with the important addition that they are then divided by the number of standard ensemble members J, in order that the sum of the weights equals one:

$$w_{t,j} = \frac{1}{J}\frac{\exp^{-t/\beta}}{\sum_{s=(0,12,24,36,48)} \exp^{-s/\beta}} \tag{2.2}$$

2.2 Description of the data and the wind farms used in the work

Before evaluating the prediction models it is important to present the wind farms and characteristics of the data used.

2.2.1 Description of the wind farms

The three wind farms studied in this work are located in different parts of France with different terrain complexity and local meteorological conditions. The sizes, in terms of number of turbines and total nominal power vary between the farms. Table 2.3 presents basic facts about the wind farms and Figure 2.2 shows their location [16]. For reasons of confidentiality, the power factor of the farms $\overline{P}/P_n$, cannot be published in this thesis. The classification of the farms according to the power factor, with the largest power factor given the lowest number, can however be given.

Mardyck and Saint Simon are rather similar in terms of terrain complexity and number of turbines. For Oupia, which has more turbines, and Saint Simon, the nominal power is identical for all turbines in each farm. Mardyck on the other hand has a configuration of three different nominal power values among the five turbines.

Table 2.3: *Facts about the wind farms.*

Farm name	**Terrain type**	**Number of turbines**	P_n **[MW]**	$\overline{P}/P_n$ **order**
Mardyck	flat	5	12.2	2
Oupia	complex	9	8.1	1
Saint Simon	flat	4	10.12	3

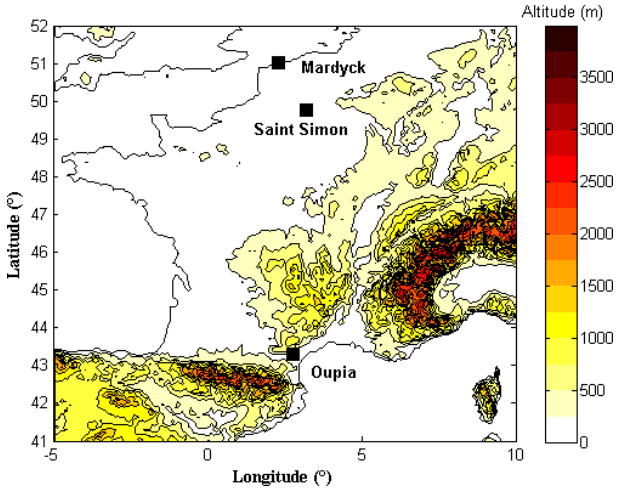


Figure 2.2: Location of the wind farms. The figure is taken from [16].

Even though the terrain complexity is similar, since Saint Simon is located in-land while Mardyck is located close to the sea, this gives different local meteorological conditions. Oupia is situated in complex terrain and its power output is therefore expected to be less predictable. The power factor, which is the average power output divided by the nominal power of the farm, is highest for Oupia and lowest for Saint Simon.

2.2.2 Characteristics of the data

The data available in this work cover the 18 months between July 2004 and December 2005. NWPs are issued twice a day, giving a total of 1098 predictions. Since power measurements for the coming days are needed to match the NWP predictions, NWP data issued for a certain date needs power measurements for the coming days ahead. Power measurements are however only given until the end of December 2005 which has for effect that the last NWP runs of the year cannot be included. In the model evaluation, the last 9 days are therefore not used giving a total of 1080 data points. These are divided into a learning set of 720 data points covering the first 2/3 of the period and a testing set of 360 data points, corresponding to the last 1/3 of the period. Since there are two NWP available per day, the learning is made over slightly less than one year, from noon 1 July 2004 to midnight 26 June 2005 while the testing is performed on almost half a year, from noon 26 July 2005 to midnight 22 December 2005.

Temporal resolution of NWP and SCADA data

The power measurements are provided by a SCADA system and consist of individual measurements for each turbine together with the total production of the farm. Measurements of meteorological variables such as wind speed and direction are provided in some of the cases. SCADA measurements are given with a temporal resolution of 10 minutes. The NWP ensemble data on the other hand are provided with a temporal resolution of 6 hours.

Since the time resolutions in the NWP and power measurements differ, some compromise

is required. In [22], where the NPRI was investigated on a Danish offshore wind farm, power measurements with a resolution of 15 minutes were used. The NWP ensemble data, which in that case also had a resolution of 6 hours, was linearly interpolated to 15 minutes. This is however not a preferable option since it is rather unlikely that this transformation is representative of the true evolution of the meteorological variables. The opposite option is to keep the 6 hour resolution from the NWP and adapt the resolution of the power measurements to that. In that case, some sort of mean of the power measurements needs to be used since the power output, as well as the wind speeds, fluctuate over shorter time scales. A mean of all power measurements over the past 6 hours would therefore not be suitable. A better way is to calculate average values for each hour and then extract the hours corresponding to the NWP data. A compromise between the two methods is also possible to obtain WPF resolutions of 1, 2 or 3 hours.

In the case of a temporal resolution of 10 minutes for the power data, obtaining the power output for hour xxh00 is therefore carried out by calculating the mean of the power measurements for 00 min, -10 min, -20 min, -30 min, -40 min and -50 min. Each sixth hour is then extracted to correspond to the NWP measurements. A resolution of 6 hours is used primarily in this work but an interpolation to a resolution of 1 hour, which is more useful in an operational context, is also carried out in order to investigate the usefulness of that option.

Spatial resolution of NWP data

The NWP data are provided on a grid of longitude - latitude pairs. The resolution is 1^o in both directions which correspond to about 75 - 80 km in East-West direction and 110 km in North-South direction. The latitude and longitude scale are included in Figure 2.2.

Since the NWP grid points do not correspond exactly to the locations of the wind farms, some kind of approximation is needed. The ideal option would be to use a complete interpolation but due to the large size of the ensembles with several dimensions, this becomes a computationally demanding alternative. Even though it might be preferable to use some kind of weighted mean of a number of grid points around the farm, since the purpose of the work is not to make as accurate forecasts as possible, it is decided to extract NWP data from the point in the NWP grid that is situated closest to the wind farm.

Both the position of the wind farms and the locations of the grid points are given in geodesic latitude and longitude. Distances between points can then be calculated with the Haversine formula which is used to calculate the distance d on Earth between two longitude - latitude pairs [1]:

$$a = \sin^2\left(\frac{\Delta_{lat}}{2}\right) + \cos\left(lat_1\right)\cos\left(lat_2\right)\sin^2\left(\frac{\Delta_{long}}{2}\right) \tag{2.3}$$

$$d = 2R\arctan\left(\frac{\sqrt{a}}{\sqrt{1-a}}\right) \tag{2.4}$$

where R is the radius of the Earth, lat_1 and lat_2 are the latitudes of the two points and $\Delta_{lat} = lat_2 - lat_1$ and $\Delta_{long} = long_2 - long_1$ with $long$ being the abbreviation for longitude.

The NWP predictions selected are provided for an altitude of 10 meter above ground level. This should be compared to the hub heights of the wind turbines for the farms which are 49, 78 and 100 m for Oupia, Saint Simon and Mardyck respectively [16]. This difference in heights means that topography, orography and other local effects have importance in the relation between the wind speeds at 10 m a.g.l and hub height. RF automatically performs the scaling to hub height but for other models such as the PC model, explicit scaling is necessary. Since Oupia is situated in complex terrain, downscaling according to Equation **??** can be expected to be less reliable than for the flat terrain wind farms Saint Simon and Mardyck.

2.3 Evaluation of the forecasting models

Before computing and evaluating the NPRI and experimenting with alternative risk indices, it is important to investigate the performance of the prediction models presented previously. In [11] a comparison between ensembles forecasts obtained using NWP ensembles from either ECMWF or the French meteorological institution ARPEGE was made. The study was carried out on the same wind farms with the same explanatory variables and on the same time period as in this case. This enables a comparison with those results. That study used primarily a probabilistic forecast model called *quantile regression forest* which is a modification of RF. The latter was however also used and comparisons with those results can therefore be made.

2.3.1 Analysis and validation of the Random Forest prediction model

The model performance is evaluated using NRMSE, NMAE and NBIAS.

Analysis of the Normalized Root Mean Square Error

Figure 2.3 shows NRMSE for Saint Simon for look-ahead times spanning the first 200 hours, which corresponds to slightly more than the first 8 days. The NRMSE is shown for the control forecast, the ensemble mean, computed with Equation 1.2, as well as the reference models persistence and climatology.

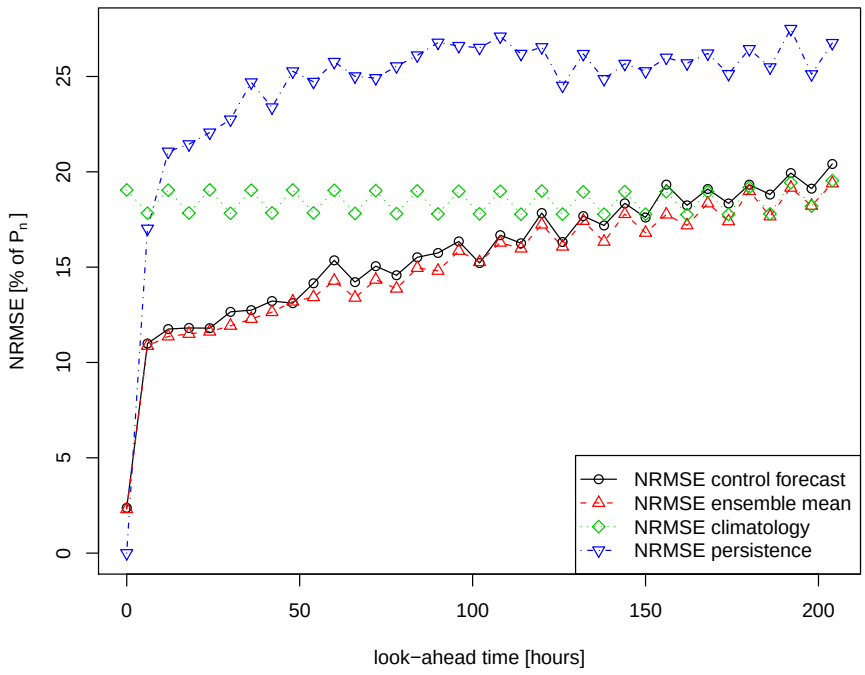


Figure 2.3: NRMSE for Saint Simon for look-ahead times from 0 to 200 hours ahead.

The forecast error is rather low for the first look-ahead time and increases then almost linearly from 6 hours onwards until it reaches the error level of climatology. This happens at about 150 hours, after which the model has the same performance as climatology or even worse. In the operational contexts considered in this work, the first 2 - 3 days are of most interest. This corresponds to the first 48 - 72 hours, which is a period in which the model performs relatively well. The inclusion of power measurements for look-ahead time 0 hours as

an explanatory variable result in a zero error for persistence for the first look-ahead time, hour 0. The persistence error then increases rapidly and stabilizes around 25 % of P_n. This shows that persistence is hard to beat for an advanced forecasting model for the first look-ahead times. The climatological error is rather constant for all look-ahead times, around 18 % of P_n. An interesting remark is that it is lower for look-ahead times 6 h, 18 h, 30 h and every 12 hours onward than for look-ahead times 0 h, 12 h, 24 h and so forth which indicates that the predictability varies diurnally.

No major improvement is found when using the mean of the ensemble members compared to the control forecast. This can probably be explained by the fact that the model is trained on the NWP control member and therefore tuned to fit the control member and not the mean of the ensembles.

For Mardyck and Oupia, the NRMSE results are presented in Figures A.1a and A.1b in Appendix A for the first 120 hours ahead. The general trend of a quasi-linear increase in NRMSE with look-ahead time is present in both those cases. The main difference compared to Saint Simon is that the error levels are higher for Mardyck and even higher for Oupia. This can be partly explained by the fact that Oupia is located in complex terrain while the other two farms are located in flat terrain. Mardyck and Oupia also have larger power factors than Saint Simon meaning that they produce proportionally more power, thus increasing the probability for larger errors. A supplementary remark to be made is that the errors for climatology and persistence vary between the farms. Persistence stays for example around 25 % of P_n for Saint Simon from 60 hours and onwards while it passes 40 % of P_n for Oupia for look-ahead times longer than 80 hours. The "zigzag" behaviour in the climatology curve that was observed for Saint Simon is present also for Oupia but not for Mardyck.

The results indicate that the RF model works well. It is however important to consider other evaluation criteria before making any final conclusions of the model performance.

Analysis of the Normalized Mean Absolute Error

The NRMSE is a useful criterion to analyze the model performance, especially when investigating the presence of large errors. However, since NRMSE is calculated as the root of the mean of the squared errors, the values of the NRMSE cannot be directly converted to a shortage or surplus of power. A criterion that is easier to interpret is the NMAE which displays the average absolute deviations, in percentage of nominal power, between the forecasted and observed power. The NMAE for look-ahead times from 0 to 120 hours ahead is shown in Figure 2.4 for Saint Simon and graphs for the other two farms are shown in Figure A.2a and A.2b in Appendix A.

For Saint Simon, the NMAE is around 8 - 9 % of P_n for look-ahead times between 10 and 40 hours. The nominal power of Saint Simon is 10.12 MW which means that in average, the predicted power will deviate by about 800 - 900 kW for these look-ahead times.

The NMAE is larger for the other two farms, with the highest values obtained again for Oupia. In all three cases, the NMAE is lower than the NRMSE, due to the fact that errors are squared in the NRMSE. The climatology NMAE varies significantly between the farms, from around 15 % of P_n for Saint Simon, 20 % of P_n for Mardyck and up to 25 % of P_n for Oupia. This order of the wind farms in terms of NMAE and NRMSE is the same as the order of the power factors, indicating that lower power factors lead to smaller errors. The zigzag behaviour is again present for Saint Simon and Oupia.

These results reinforce the trends obtained when analyzing the NRMSE. In order to get a more complete picture of the model performance, the results in terms of NBIAS are analyzed.

Analysis of the Normalized Bias

The third criterion whose use was recommended in [18] when evaluating model performance was the NBIAS. The NBIAS for the Saint Simon wind farm for look-ahead times between 0 and 120 hours ahead is shown in Figure 2.5. Graphs for the other two wind farms are shown in Figures A.3a and A.3b in Appendix A.

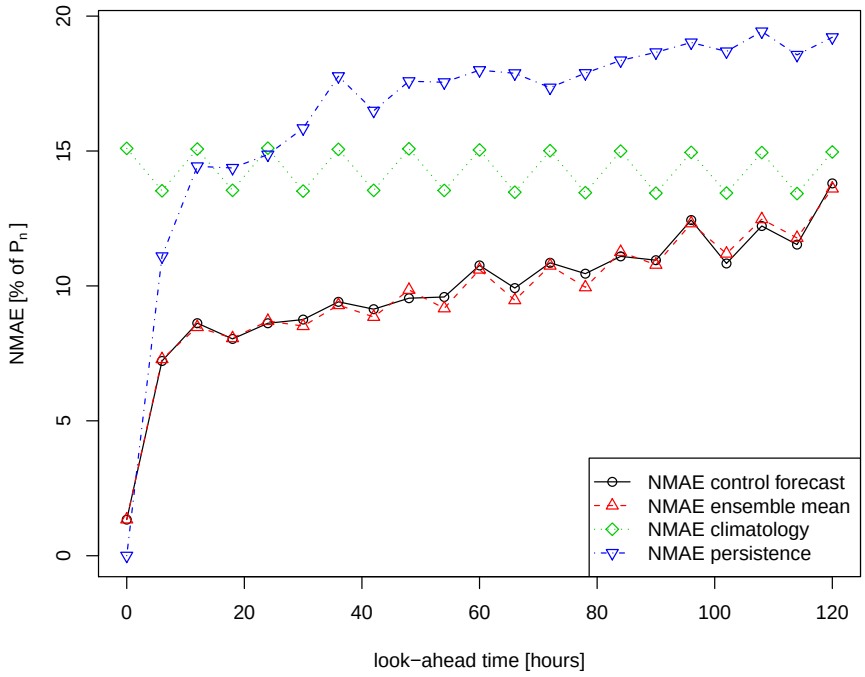


Figure 2.4: NMAE for Saint Simon for look-ahead times from 0 to 120 hours ahead.

The NBIAS varies around 0 for the control forecast and is mostly positive for the ensemble mean. The two curves show similar variations but with constantly larger values for the ensemble mean. A NBIAS closer to 0 for the control forecast is however to be expected since the model is trained on the control member and not on the mean of the NWP ensembles. For Mardyck, the NBIAS for the control forecast also varies around 0 while it is always positive for the ensemble mean. For Oupia, the NBIAS is positive for the first 50 - 60 hours and becomes negative for the rest indicating that the model performance varies with look-ahead time. The NBIAS for persistence is close to 0 which shows that the positive and negative errors that persistence makes cancels out. This is due to the variations in wind power output that goes up and down all the time. The NBIAS for climatology is positive and varies around 2 % of P_n. The zigzag behaviour found in the NRMSE and NMAE curves is also found here.

The NBIAS values for the control forecast are on average close to 0 which shows that the RF model gives a low systematic error. After studying the three criteria it can be concluded that the overall performance of RF is relatively good with results in accordance with the ones obtained in [11]. The quality of the results is thus validated.

2.3.2 Further investigation of the model performance

Even though the RF model works well, an analysis of the results when interpolating to a temporal scale of 1 hour is essential and the zigzag behaviour found requires further investigation.

Interpolation to shorter temporal scales

In an operational context, a temporal resolution of 6 hours is often too coarse. It is therefore of interest to study the model behaviour when the meteorological data is interpolated to a resolution of 1 hour. Figure 2.6 shows NRMSE and NMAE for Saint Simon with a 1 hour

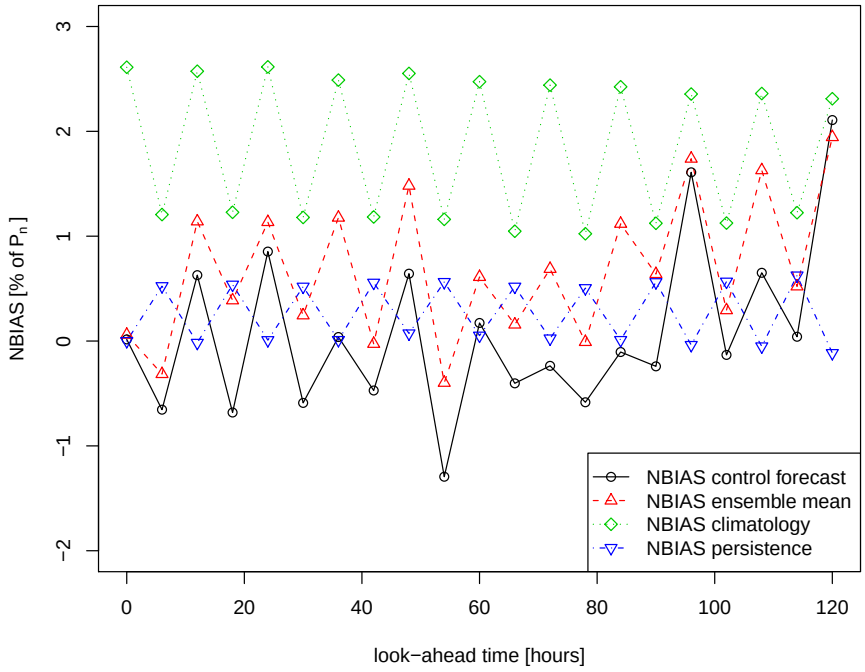


Figure 2.5: NBIAS for Saint Simon for look-ahead times from 0 to 120 hours ahead.

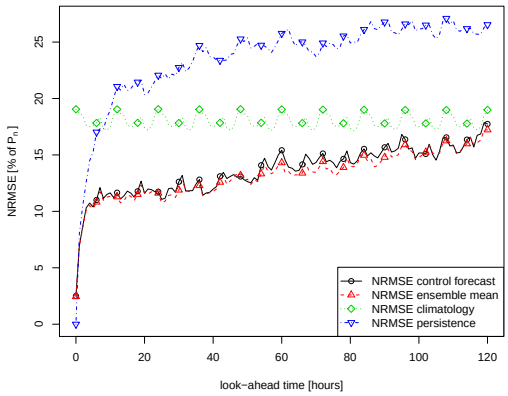
resolution for look-ahead times up to 120 hours ahead. The non-interpolated look-ahead times are marked with circles.

The general behaviour is similar to the previous results but there are small fluctuations in the error. The error is however not always increased for the interpolated look-ahead times showing that this interpolation can be performed without any problem. Since producing forecasts with shorter resolutions is more computationally demanding, a resolution of 6 hours is however used in the remainder of the work, unless stated differently.

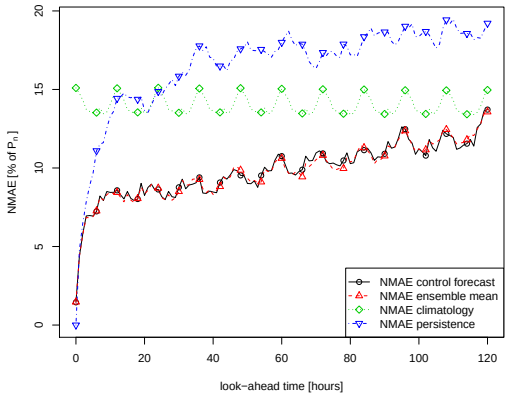
Analysis of the diurnal variations in model performance

The NWP ensemble model is run twice a day, with predictions made at midnight and noon. To investigate the differences in performance between the power forecasts based on the NWP predictions run at midnight and those run at noon, plots with the predictions separated according to run time are made. This is shown in Figure 2.7 for Saint Simon for look-ahead times up to 120 hours. Predictions made at noon are marked with circles and predictions made at midnight are marked with triangles.

Differences between the predictions made at midnight and noon are clearly observed with deviating curves. The variations are diurnal, which is especially apparent for the climatology and persistence curves for which the variations are larger. For the predictions made at noon, the errors are largest for look-ahead times 0, 24, 48 and every 24 hours onward corresponding to predictions made for the middle of the day. For the predictions made at midnight on the other hand, the errors are largest for look-ahead times of 12, 36, 60 hours and so forth, also corresponding to predictions made for the middle of the day. This indicates that the model has the most difficulty in predicting for look-ahead times in the middle of the day which is probably due to the fact that wind speeds are generally higher during the day than at night which decrease their predictability.



(a) NRMSE



(b) NMAE

Figure 2.6: NRMSE and NMAE for Saint Simon with a temporal resolution of 1 hour.

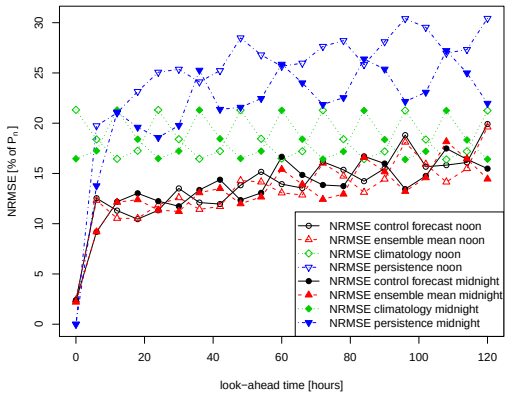
2.3.3 Analysis of the Power Curve model

The results from the PC model in terms of NRMSE and NMAE for the control forecast and the ensemble mean for Saint Simon is shown in Figure 2.8a for look-ahead times up to 120 hours ahead. For comparison, the results using the RF model and climatology are also included.

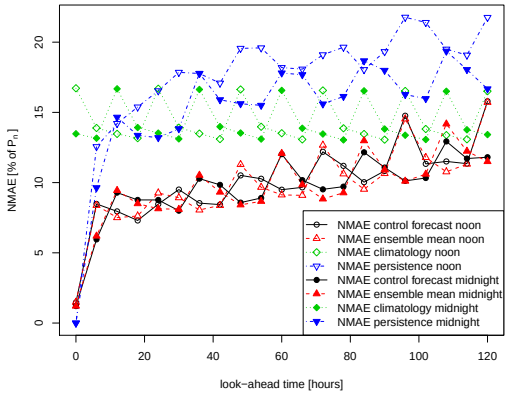
The accuracy of the predictions made with the PC model, in terms of NRMSE and NMAE, are worse than using RF, especially for look-ahead times up to 48 hours. For longer look-ahead times, the performance of the PC model is similar or even better than RF with lower values of NRMSE and NMAE compared to Figures 2.3 and 2.4 shown previously. Compared to RF, the PC model performs better in terms of NMAE than in terms of NRMSE. The PC model results show a similar zigzag behaviour as observed previously with more accurate predictions for look-ahead times 6, 18, 30 hours and so forth and worse predictions for the other look-ahead times.

Furthermore, the improvement using the ensemble mean is clearly observed. The improvement increases with look-ahead time and reaches 20 % at look-ahead time 120. This increase in improvement with look-ahead time is similar to the one found in [22]. Such an improvement was not found for RF, for which the performance of the ensemble mean and the control member was similar.

For Mardyck, the results of NRMSE and NMAE are shown in Figure A.4 in Appendix A. The results are worse than for Saint Simon with prediction errors more similar to climatology.



(a) NRMSE



(b) NMAE

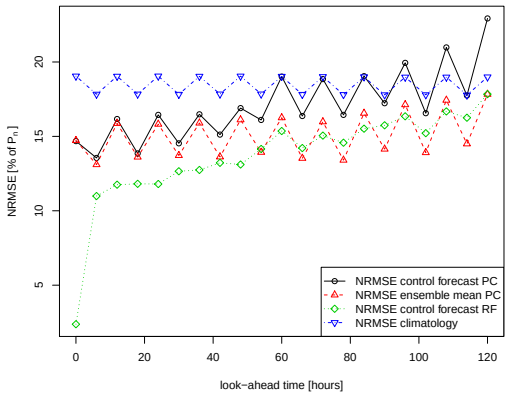
Figure 2.7: NRMSE and NMAE for predictions made at noon and midnight are separated. The plots display the results for the Saint Simon wind farm.

The control forecast from the PC model even performs worse than climatology for look-ahead times longer than 60 - 80 hours ahead. This could be explained by the fact that Mardyck is situated close to the coast and the logarithmic downscaling is less applicable for near-shore winds. The performance is even worse for Oupia, especially in terms of NRMSE for which it is outperformed by climatology. This is however not surprising since Oupia is situated in complex terrain and the simple downscaling procedure is far less accurate in such cases.

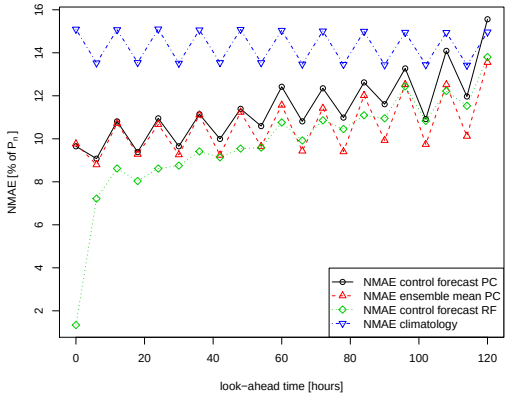
2.3.4 Improvement in NMAE and NRMSE when using a weighted mean of temporal ensembles

Concerning temporal ensembles, the possible improvement using a weighted mean of the ensemble members compared to the control forecast is evaluated. Figure 2.9 show the improvement in percentage of NRMSE and NMAE for Mardyck when using a weighted mean of the temporal ensembles obtained from the RF control members. Five weighting configurations are considered; the weights from [22], the three configurations displayed in Table 2.2 and equal weighting. Improvements are shown for look-ahead time 6 to 72 hours. The reason for not including the improvements for look-ahead time 0 is that there are large deteriorations, in the order of several hundred percent.

There are small gains in NRMSE and NMAE using temporal ensembles for the weights



(a) NRMSE



(b) NMAE

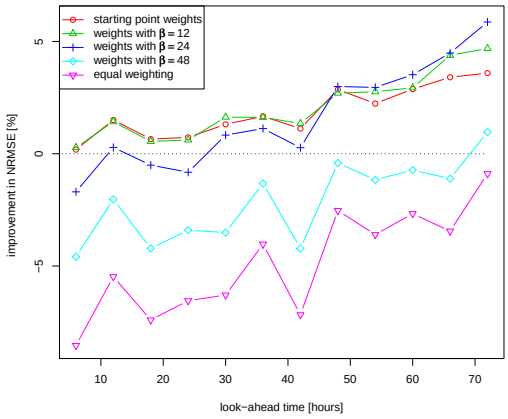
Figure 2.8: NRMSE and NMAE for the Saint Simon wind farm using the Power Curve (PC) model, Random Forest (RF) and climatology.

determined from the weights proposed in [22] and with β equal to 12. Using β equal to 24 gives a deterioration for the first 24 - 30 hours followed by an improvement for longer look-ahead times. Increasing β to 48 and equal weighting lead to worse results. There is a quasi-linear improvement with look-ahead time for all cases for NRMSE and for the first three weight configurations for NMAE. Similar results are obtained for the other two farms.

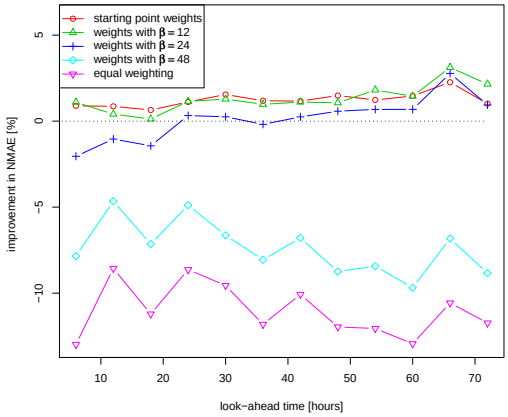
The results show that using more equal weighting of the temporal ensembles give larger errors. Certain weightings can however give lower errors than using only the latest issued forecast. Temporal ensembles can therefore be an alternative to improve prediction accuracy. The large deterioration for look-ahead time 0 is however a large drawback which indicates that this look-ahead time should be handled specifically in order to harness the full potential of this technique.

2.4 Further investigation of the models performance and characteristics of the ensembles

The evaluation of the prediction models have so far only been focused on the performance in terms of the control member. Since predictions for all ensemble members are to be used when computing risk indices it is essential that the models perform well for all members. This is



(a) NRMSE



(b) NMAE

Figure 2.9: Improvement in percentage of NRMSE and NMAE using different weightings of temporal ensembles for Mardyck.

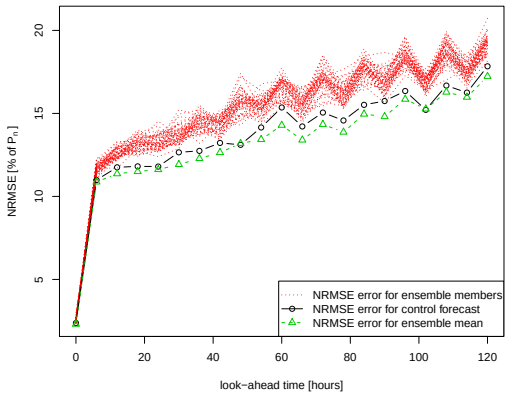
investigated by studying the average performance of the different ensemble members.

It is however not only the accuracy of the ensemble member predictions that are of interest but also the ensemble spread and especially how the spread is affected by the WPF model. Studying the ensemble behaviour for the forecast runs for which the average absolute error is the largest can help to give an indication about that. A comparison between the wind-power curve generated by the RF model and the observed relations between wind speed and power is also useful to investigate how the model fitting affect the ensemble spread. The difference in performance of the RF model and the PC model in terms of ensemble spread is then investigated using Talagrand diagrams.

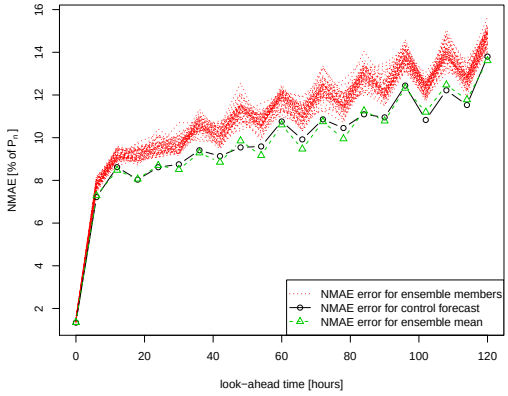
2.4.1 Analysis of the performance of different ensemble members

A way to investigate whether the ensemble members generated by the RF model have the same properties is to study the errors in terms of NRMSE and NMAE for each of the ensemble members. This is shown in Figure 2.10 for Saint Simon for look-ahead times up to 120 hours ahead. The curves for the control member are separated and error curves for the ensemble mean are also indicated.

This result confirms that the control member is the most accurate prediction among the



(a) NRMSE



(b) NMAE

Figure 2.10: NRMSE and NMAE per look-ahead time for the individual ensemble members and the ensemble mean.

ensemble members. This behaviour is however expected since the RF model is trained on the control member and the control member is assumed to be the "best guess" of the NWP ensembles. The rest of the ensemble members lie relatively close together and there is no trend that some of the ensemble members are better or worse than others. This is important since it confirms that the ensemble members have similar properties.

2.4.2 Investigation of the predictions with the largest errors

To get a better understanding of the conditions for which the RF model performs worst, the spaghetti plots for the predictions with largest total absolute error over a look-ahead time window are examined. Figure 2.11 shows spaghetti plots for look-ahead times up to 72 hours for 4 of the 8 Mardyck predictions with the largest sum of absolute errors for day 2, which is for look-ahead times between 24 and 48 hours. Vertical lines are included to illustrate the look-ahead time window on which the largest errors have been calculated.

It can be observed that the spread among the ensemble members can vary a lot over the 72 hours. It can also be seen that the spread of the ensemble members, as well as the error, can have completely different behaviour outside the period for which the largest error is calculated. An interesting remark is that when there is a large deviation between measured and predicted power, the measured power is almost always larger. This indicates that the ensemble members

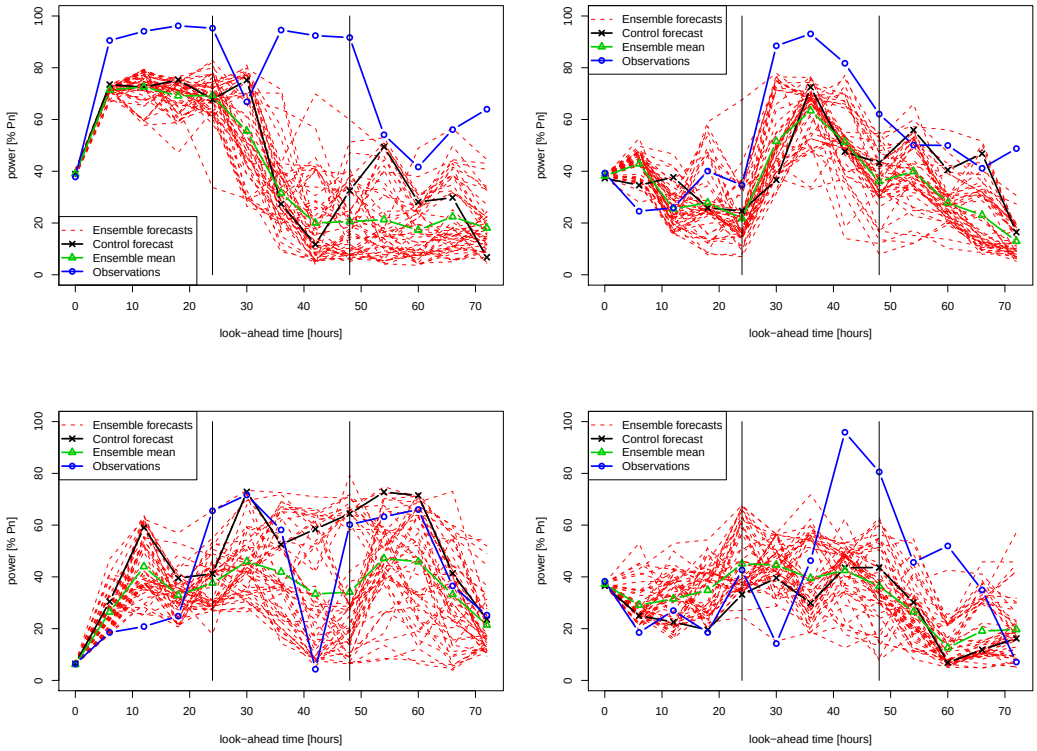


Figure 2.11: Spaghetti plots for the first 72 hours for 4 out of the 8 predictions with the largest sum of absolute errors for day 2 for the Mardyck wind farm.

are not able to forecast large power outputs.

When looking at the predictions with lowest average absolute error, these are generally found to be predictions were the power output is low or close to 0. This is to be expected since predicting is easier when the wind speeds are lower and errors in NWP predictions are not largely amplified by the WPF model. What is remarkable is that the prediction model does not predict any power value equal to 0 while 6.4 % of the power observations are equal to 0. This indicates that the ensemble members do not cover the whole range of possible power outputs. In order to further investigate this, power curves given by the prediction model are compared to the theoretical wind-power curves for the farms.

2.4.3 Difference between theoretical and modelled power curves

Power curves from the RF model are obtained by performing model learning on wind speed and direction as previously, but without taking into account the measured power at look-ahead time 0. This gives the relationship between wind speed and power for different wind directions. Artificial power curves are then obtained by testing this model on a set of manually defined couples of wind speed and direction. Wind speeds are varied between 0 and the cut-off wind speed with a resolution of 0.5 m/s and wind directions are given with a resolution of 6 degrees. Figure 2.12 shows fitted power curves together with the theoretical ones for four different look-ahead times for Saint Simon.

From the graphs, the idea that the RF model overestimates the power for low wind speeds and underestimates the power for high wind speeds is strengthen. The wind direction has a relatively large influence on the shape of the power curve, especially for wind speeds higher than 10 m/s. A grouping into a couple of main power curves can be seen, for example for look-ahead time 24 hours where two clear groups appear for wind speeds larger than 14 m/s.

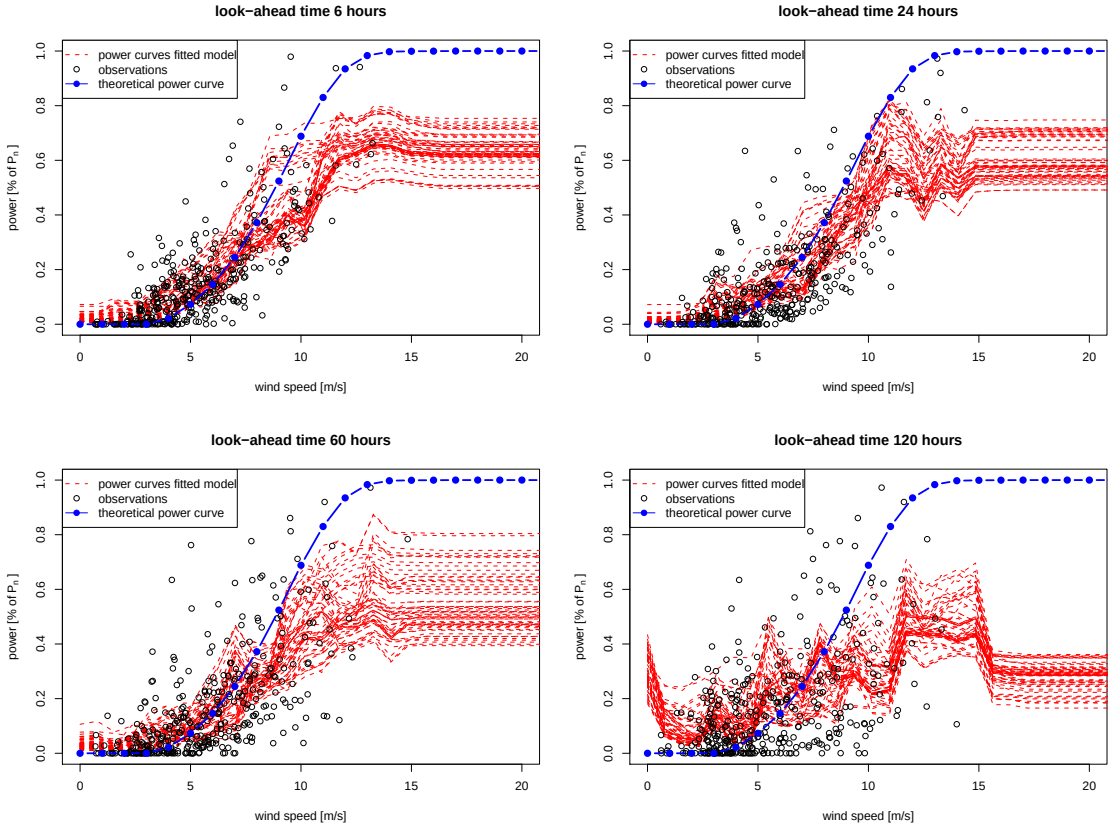


Figure 2.12: Fitted power curves for different wind speeds and directions together with observations and the theoretical power curve. The graphs are shown for Saint Simon for four different look-ahead times.

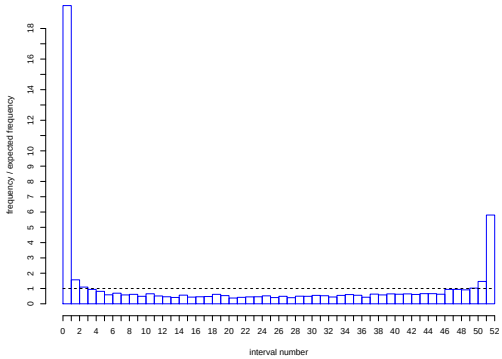
Comparing the four graphs, it is also clear that different power curves are obtained for different look-ahead times. The disagreement with the theoretical power curve increases with look-ahead time with a flattening of the fitted curves. A further remark is that the maximum power output predicted by the model does not occur for the highest wind speeds. This is probably due to the fact that few observations for larger wind speeds are present in the learning set. Another interesting remark is the zigzag behaviour in the fitted power curves with cases where the power decrease with increasing wind speeds, as can be particularly seen for look-ahead time 120 hours.

The power observations are spread out for the same wind speeds due to errors in the predicted wind speeds and the fact that all wind directions are included. The disagreement from an ideal power curve and the spread of the points increases with look-ahead time, as a result of the increasing uncertainty in the wind speed predictions. A part of this spread is however due to the fact that predicted wind speeds do not necessarily correspond to the actual ones.

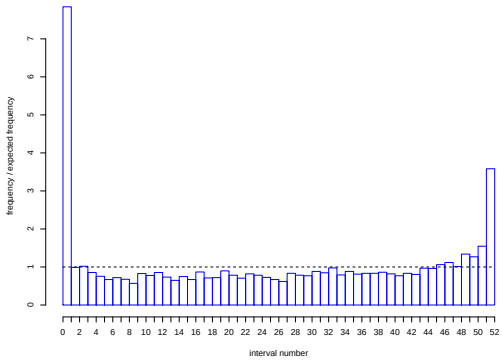
2.4.4 Analysis of ensemble member frequencies using Talagrand diagrams

The difficulty for the RF model to predict low and high power values means that there are many cases where the observed power is outside the range of predicted ensemble members. In order to verify this, Talagrand diagrams, defined in Chapter 1.4 are made for the predictions made with the RF model and the PC model. These diagrams are shown in Figure 2.13 for

the Saint Simon wind farm. Counts are shown in relative frequencies with a line included to indicate the ideal case with equal number of observations in all positions.



(a) Random Forest



(b) Power Curve model

Figure 2.13: Talagrand diagrams for the predictions made by Random Forest and the Power Curve model for the Saint Simon wind farm.

As observed, both diagrams are U-shaped showing that the observations are more often higher or lower than any of the ensemble members than they should be. This behaviour is much stronger for the predictions made by the RF model, indicating that the power ensembles generated by RF have more difficulty predicting low and high power than the PC model. This strengthens the idea that the PC model is a better option in order to have ensemble members that are more dispersed and motivates the investigation of both forecasting models when examining the NPRI.

2.5 Conclusions

In this chapter, the four methods used to produce wind power forecast ensembles in this work have been presented. Among these, multi ensembles consisting of a combination of temporal and standard ensembles have been introduced with the purpose to capture as much information as possible from the ensemble forecasts.

When evaluating the models it was found that the RF model works well. An interesting remark was the zigzag behaviour observed in the prediction error which was explained by the diurnal variations of wind speed and the fact that predictions are made either at midnight or noon. The relative difficulty in predicting for the three farms was also observed with increasing

average error with higher power factor. Furthermore it was noted that an interpolation of the NWP to a temporal resolution of 1 hour did not deteriorate the results.

In the use of temporal ensembles, no standardized way of weighting the ensembles existed previously. A method to produce such weights was therefore proposed consisting of weights that decay exponentially in time, presented by Equation 2.1. Using some of these weight configurations to calculate a weighted mean of temporal ensembles was found to give better results in terms of NRMSE and NMAE compared to using the last issued forecast. Whether these weightings also perform well when calculating risk indices is investigated in the next chapter.

Statistical prediction models such as RF are run with the objective to minimize the forecast error in terms of RMSE. This was found to have for effect that the model is conservative; it prefers to over-predict low power values and under-predict high power values, as was observed in Figure 2.12 and in the Talagrand diagram in Figure 2.13a. The PC model produces ensembles that are probabilistically more correct. This has however also for result that the prediction errors are larger. These implications are important to have in mind when calculating risk indices from the power ensembles, which is investigated in the next chapter.

Chapter 3

Analysis of the NPRI including Extension to other Temporal and Spatial Scales

After analyzing the performance of the two prediction models and the improvement obtained using temporal ensembles, it is time to make an analysis of the NPRI using the different ensembles. In [22], NPRI was evaluated on one offshore wind farm in Denmark with a nominal power equal to 5 MW, thus of the same magnitude as the three farms in this study. The NRMSE of the forecast model used in that study was similar to the one obtained for Oupia. To compare with the results obtained in that study, a similar analysis is carried out here on the three French onshore wind farms presented in Chapter 2. Since the three farms are located in different regions of France with different terrain complexities, this allows also an investigation of how the NPRI performs on different onshore terrain types.

The evaluation approach for the NPRI used in previous work is rather limited since it only covers NPRI for a specific look-ahead time window, 24 hours for either day 2 or day 3 ahead. In order to obtain a clearer picture of the performance of the NPRI depending on the position and length of the look-ahead window, different configurations are investigated in this chapter.

There can be significant variations in the forecast error for wind farms spread over an area although smoothing effects tend to reduce the total average forecast error. Providing adequate risk information at a regional scale is therefore necessary. One possible way to perform this is to calculate risk indices on the aggregate of several wind farms and another is to construct risk index maps. The extension of the NPRI to other spatial scales is therefore also covered in this chapter.

3.1 Analysis of the NPRI for day 2 or 3 for single wind farms

The NPRI is evaluated using all the four types of ensembles described in the previous chapter, starting with the power ensembles generated by the RF model.

3.1.1 NPRI derived from the Random Forest predictions

After confirming in the previous chapter that the RF WPF model performs well, investigation of the NPRI derived from its forecasts can be carried out. This is made by computing the NPRI of the ensemble members and the energy imbalances from the control member. The imbalances are normalized by the average energy imbalance $\overline{P}_{k_1}^{k_2}$ in the look-ahead time window $[k_1, k_2]$. This corresponds to the mean of the 360 energy imbalances in the testing set. The NPRI - energy imbalances pairs are divided into 5 equally populated classes and statistics of the mean,

median, upper and lower quartiles and the 10 % and 90 % quantiles are calculated for each class.

Analysis of the $NPRI_h$ for day 3

Starting with $NPRI_h$, which is the NPRI for all successive look-ahead times during one day, Figure 3.1 shows $NPRI_h$ for day 3 ahead, which is for look-ahead times between 48 and 72 hours, for the Saint Simon wind farm. The same graphs for the other two farms are shown in Figures B.1a and B.1b in Appendix B. Statistics of the mean and the quantiles in each class are displayed at the average NPRI value in each class.

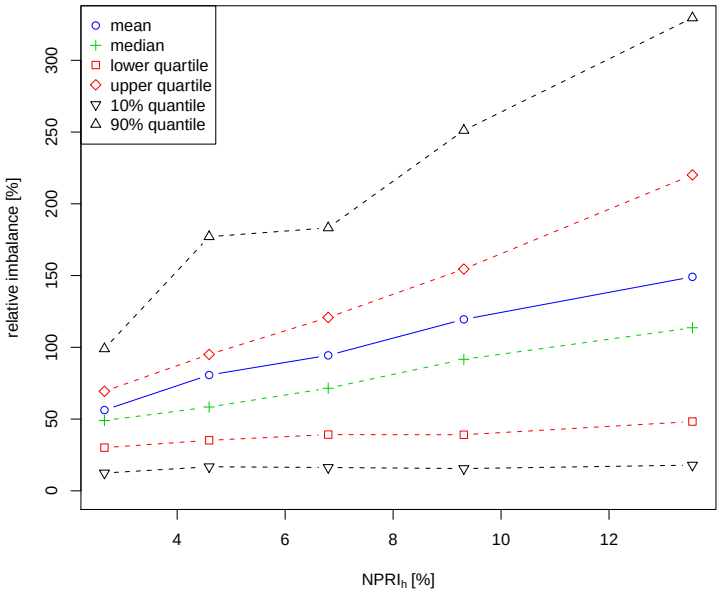


Figure 3.1: $NPRI_h$ for day 3 for Saint Simon.

The general behaviour is similar to the one obtained in [22] with a mean relative energy imbalance that increases almost linearly with look-ahead time. There are however two important differences compared to those results. The first difference is that the NPRI values are lower for these farms. This indicates that the ensemble members lie closer together since lower spread of ensemble members leads to a lower weighted standard deviation and thus to a lower NPRI. The second difference is that the ratio of energy imbalance in the fifth and first class is smaller. The ratio is equal to 2.67 for Saint Simon and values of 2.45 and 1.97 are obtained for Mardyck and Oupia. This ratio was equal to 5.4 for the offshore wind farm presented in [22]. This reduction is mainly due to a higher relative mean imbalance level for the first class. These smaller differences in energy imbalances between the classes indicate that the $NPRI_h$ does not discriminate as well between situations with high and low energy imbalance as in the Danish offshore case. The mean of the relative imbalance level is higher than the median for all three cases, a trend that was also observed in [22].

The sharpness of the relative imbalance distributions is measured by the IQR, defined in chapter 1.6. It can be used to measure the skill of the index which is defined as the ability to predict the imbalance level. The IQR ranges from 41 % in the first class to 167 % in the last for Saint Simon and these values are equal to 43 - 160 and 58 - 136 % for Mardyck and Oupia respectively. There is thus a growth in IQR with class number, as can be observed in Figure 3.1. This growth indicates that there is a larger spread of the energy imbalances for larger NPRI values. Since the energy imbalances are in average larger for higher classes, this implies

also a larger risk for large imbalances. Even though the resolutions are lower than in [22], the $NPRI_h$ can therefore still be useful to inform on the risk for large imbalances.

In [22], the $NPRI_d$ was found to give sharper energy imbalance distributions and it is now interesting to examine whether that is the case here.

Analysis of the $NPRI_d$ for day 3

When the NPRI is calculated for look-ahead time windows, the distribution of relative imbalances for each class becomes narrower which is well illustrated in Figure 3.2. Figures for the other two farms are shown in Figure B.2a and B.2b in Appendix B.

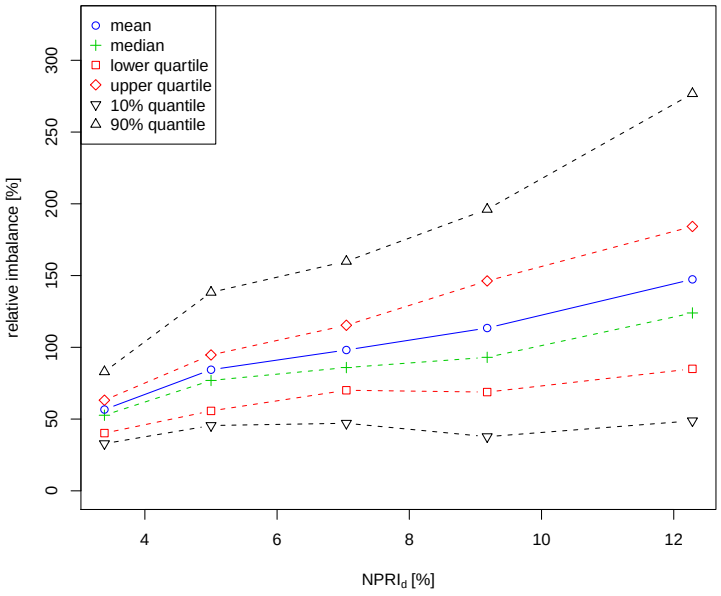


Figure 3.2: $NPRI_d$ for day 3 for Saint Simon.

These results are also similar to the ones observed in [22]. The ratio of relative imbalance level in class 5 and class 1 are however again smaller, equal to 2.50 compared to 4.2 in the Danish offshore case. Even smaller values are obtained for the other two farms, 2.08 for Mardyck and 2.01 for Oupia. Compared to $NPRI_h$, there are reductions for Saint Simon and Mardyck while there is actually an increase in this ratio for Oupia. The values of the imbalance levels in the first classes are relatively high, more than 50 % of the average imbalance already in the first class. This means that even when the spread of the ensemble members is lowest, giving low $NPRI_d$ values, the energy imbalance is in average still greater than 50 % of usual. It would be preferable to obtain a larger difference in relative imbalance between the classes since this would mean that the index can better distinguish between more and less predictable situations. The plots are however still informative since it is clearly observed that the risk for larger energy imbalances increases with $NPRI_d$ value. For example, the energy imbalance for Saint Simon is never above average when the $NPRI_d$ is in the first class while it can be expected to be larger than usual when the $NPRI_d$ is in the last class.

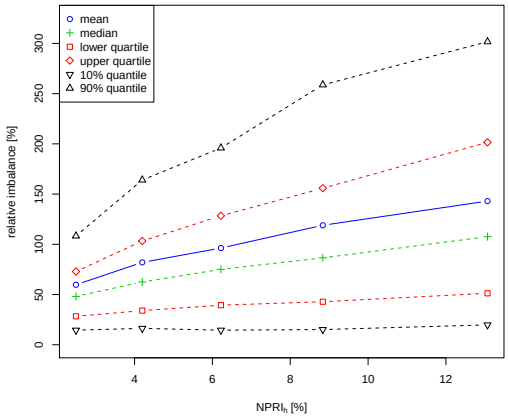
The narrower distributions are, as mentioned above, preferable in a skill forecasting point of view. The ranges of IQR from class 1 to class 5 are equal to 23 - 97 % for Saint Simon, 33 - 69 % for Mardyck and 35 - 71 % for Oupia. The reduction in IQR is present for all classes but is most pronounced for the higher ones. The IQR is for example reduced by more than half for class 5 for Oupia. Even though the distributions are narrower, there is still a large difference between the different classes. In the case of Saint Simon, for which the differences between the

classes is largest, the IQR is relatively low for the first class meaning that an $NPRI_d$ value in that class gives a rather high certainty of the relative imbalance level compared to an $NPRI_d$ in the fifth class where the spread is relatively large.

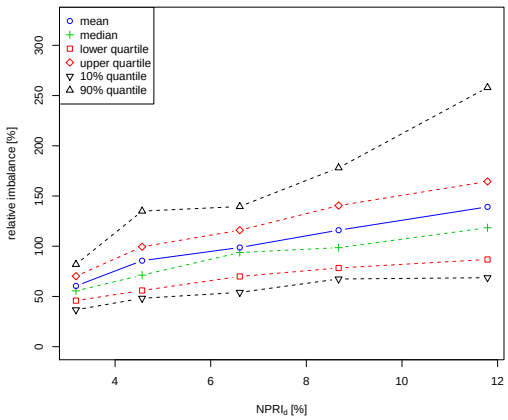
The values of the upper quantiles are also of interest since they provide information on the probability or risk for large energy imbalances. It can for example be seen in Figure 3.2 that there is about 10 % risk for an energy imbalance twice as large as normal when the $NPRI_d$ is found in class 4. The risk for such a large relative energy imbalance is very low for lower classes while it is even higher when $NPRI_d$ lies in class 5.

NPRI using forecasts interpolated to shorter temporal scales

In [22], the NPRI was evaluated with predictions made on a temporal scale of 15 minutes. In the previous chapter it was found that interpolating NWP to a resolution of 1 hour did not deteriorate the forecast performance in terms of NRMSE and NMAE. The implications of the interpolation when producing risk indices is evaluate here. Figure 3.3 show $NPRI_h$ and $NPRI_d$ for day 3 for Saint Simon, derived from wind power ensembles made with a 1 hour resolution.



(a) $NPRI_h$



(b) $NPRI_d$

Figure 3.3: $NPRI_h$ and $NPRI_d$ for day 3 for Saint Simon with a 1 hour resolution.

The ratios of mean imbalance level in class 5 and class 1 are here equal to 2.37 for $NPRI_h$ and 2.30 for $NPRI_d$. These ratios are thus slightly lower than for the calculation on non-

interpolated data. A small reduction in the NPRI values can also be observed. Furthermore, the IQRs vary from 43 - 151 % for $NPRI_h$ and 25 - 76 % for $NPRI_d$ showing that the distributions are narrower, especially for larger NPRI values. It can particularly be observed that all quantiles increase with class number for $NPRI_d$, which was not the case when using a temporal resolution of 6 hours, shown in Figure 3.2

Albeit the slightly reduced ratios of mean imbalance level, the narrower distributions are preferable since they give a sharper estimation of the relative error. Calculating NPRI on predictions interpolated to a 1 hour resolution can therefore be made without any problem. This is an important result which confirms that this method can be used in an operational context even when the temporal resolution of the NWP forecasts is coarser than 1 hour.

Comments on the results

At a first glance, the results obtained in this work can seem to be disappointing compared to those found in [22]. Lower NPRI values are found indicating that the dispersion among the power ensemble members is lower in the cases in this work. Furthermore, the lower ratios of imbalance level between classes 5 and 1 mean that the NPRI is less useful in distinguishing between situations with high and low energy imbalances. In terms of sharpness of the distributions of relative imbalance, the NPRI is however rather fine in these cases, especially for $NPRI_d$ where it performs as well as in [22]. The larger IQR values for $NPRI_h$ than $NPRI_d$ also strengthens the idea that the $NPRI_d$ is more efficient in determining the relative imbalance level depending on the NPRI value. The risk indices can therefore still be informative, especially in informing on the risks for large imbalances.

Notable is also the fact that ratios of mean imbalance level in class 5 and class 1 are even lower compared to when temporal ensembles were used for the Danish wind farm in [22]. The usefulness of temporal ensembles in this case is therefore interesting to examine.

3.1.2 NPRI derived from temporal ensembles

In the preceding chapter it was shown that some of the weight configurations in Table 2.2 employed to calculate a weighted mean of temporal ensembles improved forecast accuracy compared to the last issued forecast. Whether the same weight configurations performs well when using them to calculate the weighted standard deviation to make NPRIs is examined here. Statistics for $NPRI_h$ and $NPRI_d$ using the weight configurations with β equal to 12 and 24 are shown in Table 3.1 for the three farms. Three criteria are compared; the ranges of NPRI displayed as the mean NPRI in class 1 and 5, the ratios of mean relative energy imbalance in class 5 and 1, and the minimum and maximum IQR among the classes. The minimum and maximum IQR are almost always found for classes 1 and 5 respectively. Exceptions are however found, which is the reason why the minimum and maximum IQR are shown and not the IQR for class 1 and 5.

The first three rows of the table contain the average NPRI for class 1 and 5 for the three farms. The actual NPRI values are however not important, since it is rather a values position among the distribution of NPRI values that is of interest. Looking at the ranges of NPRI values is interesting nevertheless, since it informs on the spread of ensemble members, and thus gives a relative indication of how much the members are spread for different types of ensembles and for different farms. This shows that the NPRI values are smallest for Saint Simon and largest for Oupia. This means that the spread of the ensemble members is generally largest for Oupia. With respect to this criterion, the farms are ordered in the same order as their power factor. It can also be observed that the NPRI values are lower than when using standard power ensembles.

The ratio of mean imbalance level in the fifth and the first class is displayed in rows 4 to 6. These values are also lower than the ones found using standard power ensembles. The performance in terms of ratio of mean imbalance level in class 5 and class 1 shows that the

Table 3.1: NPRI results using temporal ensembles.

	Wind farm	$NPRI_h$ [%] β = **12**	$NPRI_d$ [%] β = **12**	$NPRI_h$ [%] β = **24**	$NPRI_d$ [%] β = **24**
mean NPRI class 1-5	Mardyck	2.3 - 13.4	3.4 - 11.4	2.5 - 13.1	3.6 - 11.3
	Oupia	2.7 - 17.1	4.0 - 14.6	3.1 - 16.4	4.3 - 14.3
	Saint Simon	1.8 - 11.3	2.7 - 9.7	2.1 - 11.1	2.9 - 9.7
ratio of mean imbalance level	Mardyck	1.93	1.84	2.01	1.87
	Oupia	1.43	1.56	1.44	1.62
	Saint Simon	1.92	2.04	1.94	2.03
minimum and maximum IQR	Mardyck	54 - 166	38 - 66	52 - 165	34 - 63
	Oupia	73 - 136	44 - 72	64 - 126	43 - 69
	Saint Simon	47 - 154	32 - 91	46 - 162	31 - 91

indices are still rather informative with values around 2, except for Oupia for which the values are lower. Varying the weights does not have a large effect on the index performance. Similar results as in Table 3.1 are obtained when β is set to 48 while worse results are obtained when using the weights found in [22]. Using equal weighting gives even worse results.

Rows 7 to 9 display the minimum and maximum IQR among the five classes. These values are larger than those found using standard power ensembles, which shows that temporal ensembles provide less useful information on the relative imbalance than standard ensembles.

Even though the results are worse than those found using standard power ensembles, the NPRI from temporal ensembles can still distinguish between energy imbalance levels to some extent. They can thus be a useful alternative to ordinary power ensembles, especially since they are simple and cheap to obtain from single wind power forecasts. Concerning the choice of weights, the results indicates that some of the weight configurations presented in Table 2.2 are more appropriate to use than the weights found in [22] or equal weighting. It is therefore interesting to investigate whether the same trends are found when using multi ensembles.

3.1.3 NPRI derived from multi ensembles

Multi ensembles are derived by making temporal ensembles on all standard ensemble members giving a larger set of members. Statistics when the NPRI is calculated on these ensembles are shown in Table 3.2 using weights from Equation 2.2 with β equal to 12 and 24.

Table 3.2: NPRI results using multi ensembles.

	Wind farm	$NPRI_h$ [%] β = **12**	$NPRI_d$ [%] β = **12**	$NPRI_h$ [%] β = **24**	$NPRI_d$ [%] β = **24**
mean NPRI class 1-5	Mardyck	5.3 - 16.5	5.9 - 15.5	5.6 - 16.2	6.1 - 15.4
	Oupia	7.1 - 19.5	7.9 - 18.4	7.5 - 19.3	8.2 - 18.5
	Saint Simon	4.1 - 14.2	4.7 - 13.0	4.4 - 13.9	4.9 - 12.8
ratio of mean imbalance level	Mardyck	2.31	2.20	2.25	2.23
	Oupia	1.69	1.84	1.62	1.77
	Saint Simon	2.41	2.46	2.40	2.33
minimum and maximum IQR	Mardyck	52 - 169	38 - 64	53 - 164	38 - 69
	Oupia	55 - 122	37 - 65	59 - 125	38 - 65
	Saint Simon	42 - 174	27 - 95	43 - 168	32 - 95

Comparing with Table 3.1, a better performance than the ordinary temporal ensembles is obtained. There is however no improvement compared to the NPRI calculated on the standard

power ensemble members, except for Mardyck $NPRI_d$. The NPRI ranges are increased for both $NPRI_h$ and $NPRI_d$ for all three farms compared to standard ensembles. Again, no major differences between different weightings are found, except for the weights proposed in [22] and the equal weighting which both give worse results than the weights determined by β equal to 12 or 24.

Since this method to calculate NPRI does not improve over NPRI calculated from standard ensembles, there seems to be no usefulness of using such ensembles.

3.1.4 NPRI derived from the Power Curve model ensemble predictions

The lesser results for the NPRI in this work are probably due to the fact that the meteorological ensembles are not correct in a probabilistic sense. Using a prediction model like RF that does not preserve the ensemble spread increases this problem. Since it was shown in the Talagrand diagrams in Figure 2.13 in Chapter 2 that the PC model better preserve the ensemble spread it is well motivated to evaluate the NPRI on the predictions made by this model.

NPRI results when calculating both NPRI and energy imbalances from the Power Curve model predictions

Results of NPRI using the PC model are shown in Figure 3.4 for day 3 for Saint Simon.

The ratio of mean imbalance level in class 5 and 1 is far higher than using RF, equal to 5.72 and 5.15 for $NPRI_h$ and $NPRI_d$ respectively. This shows that there is a better distinction between situations with high and low energy imbalance. The energy imbalances are however larger, as it was observed in Figure 2.8 in Chapter 2 with forecast larger forecast errors using the PC model than the RF model for the considered look-ahead time window.

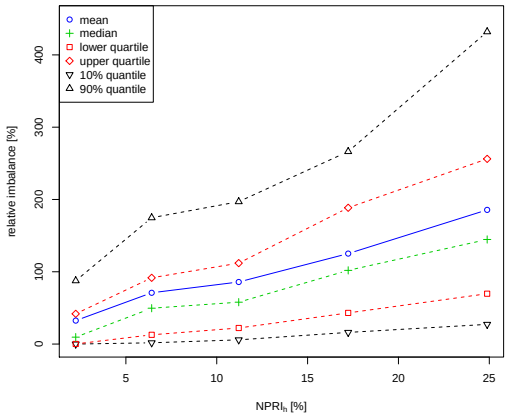
The next remark to be made is that the NPRI values are significantly larger than using the RF model. This shows that the PC model gives a larger spread of the ensemble members for higher power outputs and a lower spread for lower power outputs. The IQRs of the energy imbalance distributions are similar as for the RF results. The results for Mardyck and Oupia also show improvements in terms of larger ratio of mean imbalance level, compared to the results found using RF.

The higher resolution in NPRI from the PC model means that it gives better information on the size of expected energy imbalance depending on the ensemble spread. However, since the energy imbalances are larger, these predictions are less useful if the objective is to produce the most accurate forecasts as possible in terms of NMAE or NRMSE. The question is then whether it is preferable to produce lower errors with less information about the errors or larger errors but with more information about the errors. The ideal case would be to have a model that gives low errors but also distinguishes well between more and less predictable situations. It is therefore interesting to study whether there is a strong relation between the energy imbalances obtained from the predictions made by the RF model and the NPRI obtained using the PC model predictions.

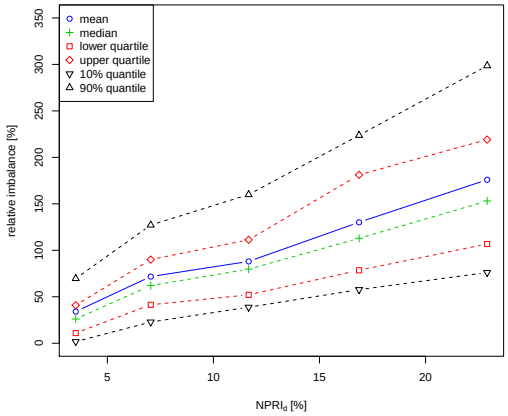
NPRI results when calculating the NPRI from the Power Curve model predictions and energy imbalances from the Random Forest model predictions

The $NPRI_d$ plot when using the PC model ensemble predictions to calculate the NPRI and using the RF model control forecast to calculate the energy imbalances is shown in Figure 3.5a for day 2 for Saint Simon. The corresponding graph when using both NPRI and energy imbalances on the RF predictions is included in Figure 3.5b for comparison.

Even though some minor differences between the two graphs are observed, the performance is very similar with almost equal ratios of mean imbalance level, 2.39 and 2.46 respectively, and similar IQR, 29 - 89 % and 28 - 96 %. The main difference is that the $NPRI_d$ values are larger in Figure 3.5b. This is however to be expected since the $NPRI_d$ is calculated on the predictions made with the PC model which has a larger spread.



(a) $NPRI_h$



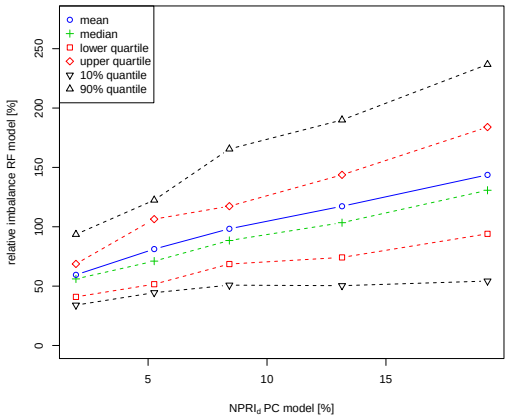
(b) $NPRI_d$

Figure 3.4: $NPRI_h$ and $NPRI_d$ for day 3 for Saint Simon from predictions made with the Power Curve model.

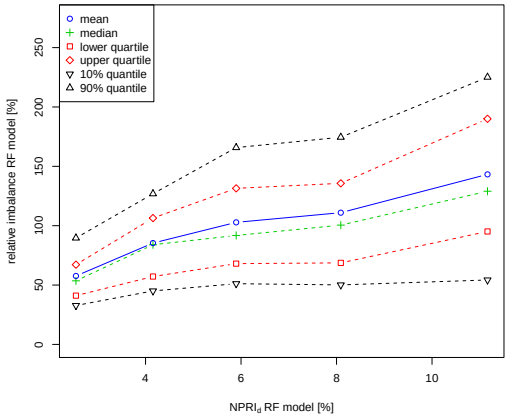
The results for Mardyck and Oupia are displayed in Table 3.3 which shows statistics for $NPRI_d$ for day 2 for the three wind farms when using the RF model and the PC model with energy imbalances from the RF model control forecast. The $NPRI_d$ calculated on the temporal ensembles and multi ensembles from the PC model, still using energy imbalances from the RF model, are also included. Oupia is included even though the downscaling of wind speeds is not reliable.

Using the standard ensembles from the PC model is shown to give better results than temporal ensembles or multi temporal ensembles from the same model. The reduction in ratio of mean imbalance level using the PC model is larger for Oupia than for Mardyck and Saint Simon. The IQR ranges are rather similar for the different options. The lowest values for the first class are however found using RF.

Calculating the $NPRI_d$ on temporal ensembles obtained from the PC model also gives better results than the $NPRI_d$ for temporal ensembles made by RF. In terms of ratio of mean imbalance level in class 5 and 1, the results are even better than the $NPRI_d$ calculated on the standard power ensembles from RF. The $NPRI_d$ from the temporal ensembles does not



(a) $NPRI_d$ from PC



(b) $NPRI_d$ from RF

Figure 3.5: $NPRI_d$ for day 2 for Saint Simon when the NPRI is calculated on the ensemble predictions made with the PC model (a) and the RF model (b). The energy imbalances are obtained from the RF model control forecast in both cases.

Table 3.3: NPRI results using the standard ensembles, temporal ensembles and multi ensembles from the Power Curve (PC) model compared to the standard Random Forest (RF) ensembles. The energy imbalances have been calculated on the standard RF control forecast.

	Wind farm	$NPRI_d$ **RF stand.**	$NPRI_d$ **PC stand.**	$NPRI_d$ **PC temp.**	$NPRI_d$ **PC multi**
mean NPRI class 1-5	Mardyck	2.6 - 13.2	2.8 - 21.8	1.3 - 17.1	5.0 - 25.2
	Oupia	5.6 - 14.8	1.0 - 18.3	0.1 - 8.7	1.5 - 19.9
	Saint Simon	2.5 - 11.1	1.9 - 19.1	0.7 - 13.6	3.1 - 21.6
ratio of mean imbalance level	Mardyck	2.67	2.57	2.22	2.37
	Oupia	2.01	1.76	1.65	1.74
	Saint Simon	2.49	2.39	2.40	2.33
minimum and maximum IQR	Mardyck	23 - 81	24 - 94	24 - 82	23 - 79
	Oupia	27 - 65	30 - 62	45 - 72	29 - 62
	Saint Simon	25 - 91	28 - 85	31 - 97	40 - 99

however give better results than the corresponding power ensembles with the same model. Neither are the results improved when using multi ensembles obtained from the PC model ensemble predictions.

Correlations between energy imbalances and the NPRIs obtained from the two prediction models

Even though the combination of NPRI from the PC model and energy imbalances from the RF model does not consistently improve NPRI performance compared to using only the RF model it is still interesting to further investigate this option. Scatter plots of the relations between the $NPRI_d$ calculated from the two indices and the energy imbalances from the RF model control forecast are therefore made. These plots together with the correlation coefficients are shown in Figure 3.6 for day 2 for Saint Simon.

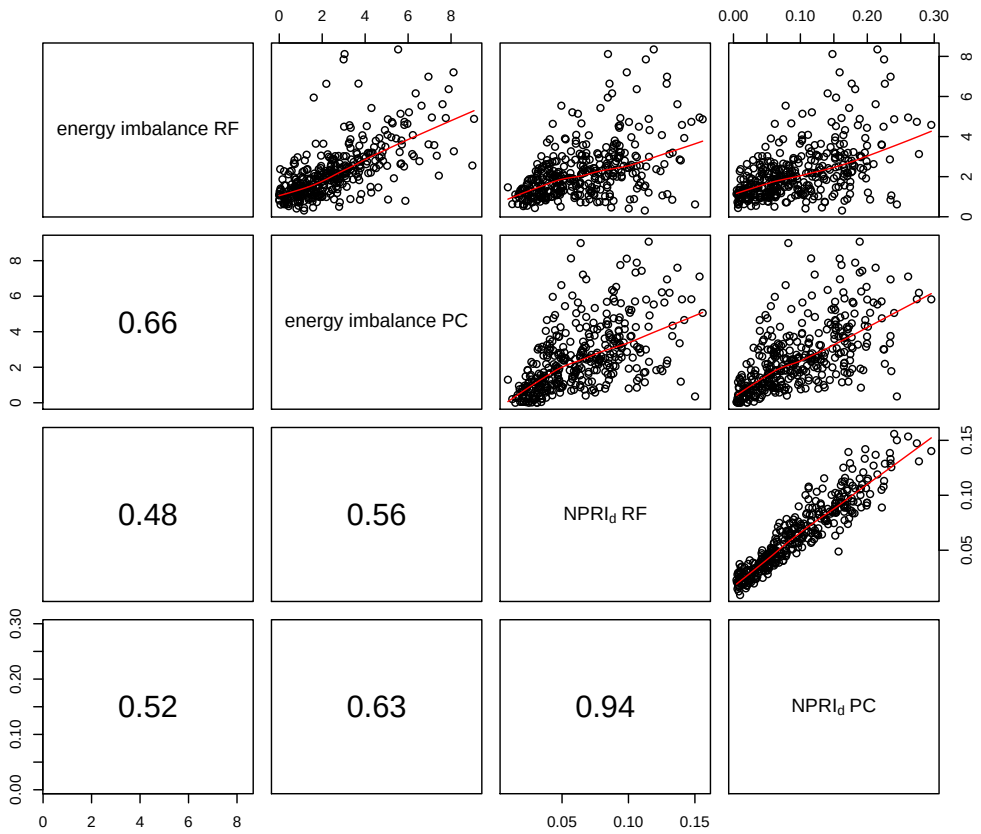


Figure 3.6: Correlations between the energy imbalances from the Random Forest (RF) model and the $NPRI_d$ for the ensemble predictions made by the RF model and the Power Curve (PC) model. The results are for day 2 for the Saint Simon wind farm.

There are a number of large deviations from a linear relation between the energy imbalances from the two models with a correlation coefficient equal to 0.66. This stems from the fact that the predictions from the two models are not identical and the energy imbalances are therefore not the same using the two models. The results also show a rather strong correlation between the $NPRI_d$ calculated from the ensembles produced by the two models.

Comparing the correlations between the $NPRI_d$ from RF and the $NPRI_d$ from the PC model with the energy imbalances obtained using RF shows that the correlation with the PC model $NPRI_d$ is stronger, indicating that it better explains the energy imbalances. The

difference is however rather small with a correlation of 0.52 compared to 0.48. There seem thus not to be a large gain in using the PC predictions as a complement to the RF predictions. In the remainder of this report, all predictions are therefore made with the RF model if not stated differently.

3.2 Extension of the NPRI to other temporal scales

So far, the look-ahead time window employed has had a length of 24 hours. In the spaghetti plots presented in Figure 2.11 in Chapter 2 it was observed that the spread of the ensemble members can vary significantly over a period of 2 - 3 days. This motivates the examination of how the NPRI performs when the look-ahead time window is longer or shorter by regarding a doubled or halved time window. The location of the look-ahead time window is also of interest as well as the evolvement of the ensemble spread in time.

3.2.1 Evolvement of the spread of the ensemble members with look-ahead time

When using the same weights for all ensemble members, Equation 1.1 is simplified to:

$$\widetilde{\sigma}_{t,k} = \left[\frac{1}{J-1} \sum_{j=1}^{J} \left(\widehat{P}^{(j)}_{t+k|t} - \overline{P}^{J}_{t+k|t} \right)^2 \right]^{1/2} \tag{3.1}$$

which is the expression for the estimate of the standard deviation of the ensemble members. In the case of $k_1 = k_2$, the above expression is also the expression for the $NPRI_h$.

Evolution of the spread of the power ensembles with look-ahead time

Figure 3.7 shows the average $NPRI_h$ values for each look-ahead time together with the median, the upper and lower quartile and the 10 % and 90 % quantiles for Saint Simon. The results are obtained from the same data set used previously, giving statistics out of 360 predictions. Results are shown for non-interpolated data for look-ahead times covering the first 9 days ahead. Graphs for the other two farms are shown in Figure B.3 in Appendix B.

The average standard deviation and all quantiles increase with look-ahead time for the first 90 hours after which they decline. The spread, given by the distances between the quantiles, also increases slightly up to 90 hours after which it also decreases. The graphs are similar for the other two farms with the exception that the standard deviation of the ensemble members is highest for Oupia followed by Mardyck and Saint Simon for all look-ahead times. This trend was also found when looking at the ranges of NPRI values. Notable is also that the mean and median values are similar.

The decrease in ensemble spread is not expected. A possible explanation could be that the fitted power curves obtained with the RF model get flatter with look-ahead, as observed in Figure 2.12 in Chapter 2. Whether the decrease in ensemble spread for longer look-ahead times are due to the WPF model or stems from the NWP ensembles is therefore a subject for investigation.

Evolution of the spread of the wind speed ensembles with look-ahead time

The same graphs are therefore made with the $NPRI_h$ calculated on the wind speed ensembles. Results are shown in Figure 3.8 for Saint Simon and in Figure B.4 in Appendix B for the other two farms. The maximal wind speed during the evaluation period is used as a normalization factor.

Contrary to the power ensembles, the standard deviation of the wind speed ensembles does not decrease for longer look-ahead times. Only a small stagnation is found after about

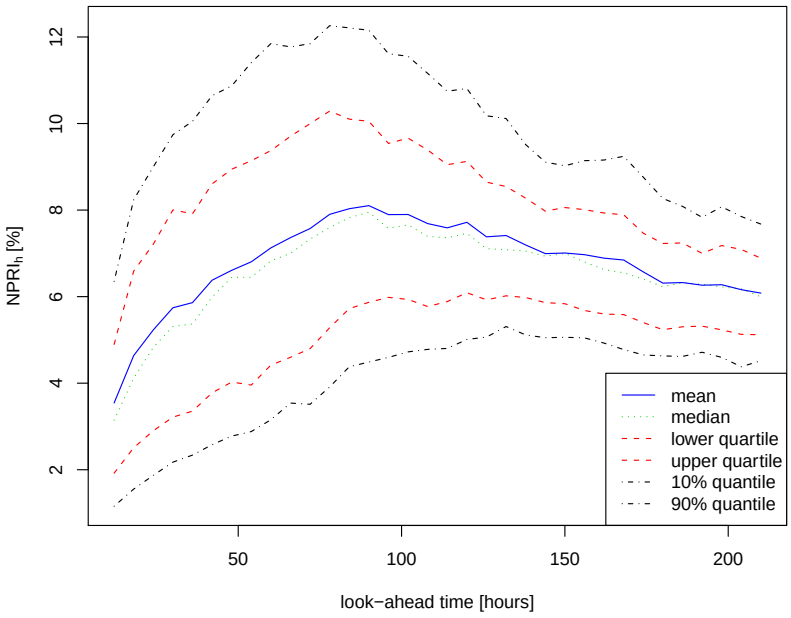


Figure 3.7: $NPRI_h$ per look-ahead time for Saint Simon for look-ahead times covering the first 9 days ahead.

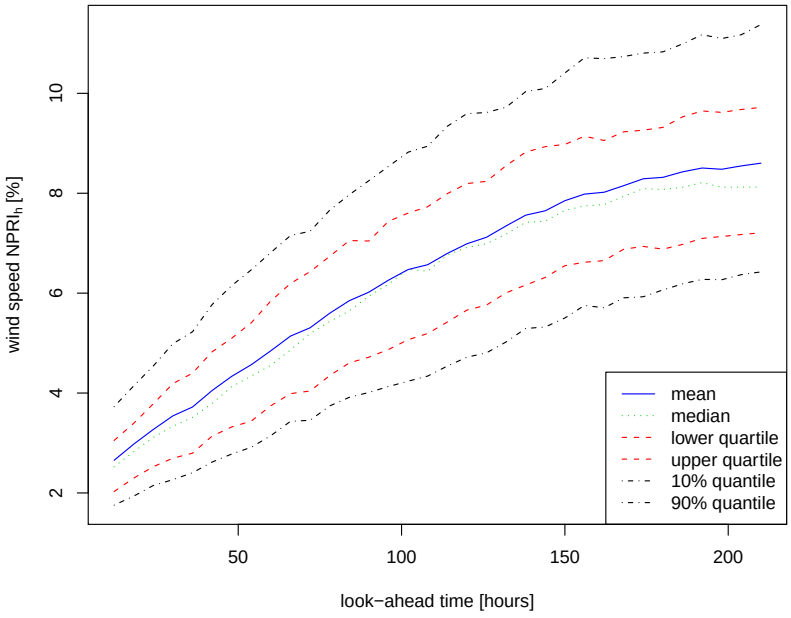


Figure 3.8: $NPRI_h$ of the predicted wind speed ensemble members per look-ahead time for Saint Simon for look-ahead times covering the first 9 days ahead.

150 hours. The variation of the standard deviations also increases with time, contrary to the behaviour found for the power ensembles. Similar behaviour is found for the other two farms.

The increased spread of the wind speed ensembles with time is due to decreased predictability. The patterns in the figures are therefore not surprising and confirm that the meteorological ensembles function well in the sense that the dispersion among the alternative trajectories increases with look-ahead time. The disagreement in the evolution of the wind speed ensemble spread and the power ensemble spread must therefore be explained by the wind-power conversion. Since, the fitted power curve becomes flatter for longer look-ahead times, as observed in Figure 2.12 in Chapter 2, this decreases the ensemble spread for longer look-ahead times. The spread of the power ensembles is therefore most representative for the first 4 - 5 days and considering longer look-ahead times can be expected to give less relevant information in the purpose of risk indices. The differences during the first 5 days is however interest to investigate.

3.2.2 NPRI for different days

Up until now the NPRI and energy imbalances have been calculated for look-ahead times corresponding to day 2 or day 3 ahead. As the ensemble spread is rather wide for the first 4 - 5 days ahead, as observed in Figure 3.7, it is of interest to investigate the results of the $NPRI_d$ for different days in this time interval. Plots of $NPRI_h$ and $NPRI_d$ for day 1 are shown for Saint Simon in Figure B.5 in Appendix B and statistics of the results when calculating $NPRI_d$ on days 1, 2, 3, 4 and 5 are shown in Table 3.4 for the three farms.

Table 3.4: $NPRI_d$ results for different days.

	Wind farm	$NPRI_d$ **Day 1**	$NPRI_d$ **Day 2**	$NPRI_d$ **Day 3**	$NPRI_d$ **Day 4**	$NPRI_d$ **Day 5**
mean NPRI class 1-5	Mardyck	1.7 - 9.5	2.7 - 13.4	4.2 - 14.7	5.2 - 15.3	5.9 - 14.6
	Oupia	4.4 - 13.2	5.6 - 15.0	6.6 - 17.2	7.3 - 19.0	7.6 - 18.1
	Saint Simon	1.6 - 8.6	2.5 - 11.2	3.4 - 12.3	4.3 - 12.5	4.7 - 11.0
ratio of mean imbalance level	Mardyck	3.14	2.65	2.08	2.06	1.69
	Oupia	1.96	2.00	2.01	1.80	1.51
	Saint Simon	3.01	2.44	2.50	1.99	1.72
minimum and maximum IQR	Mardyck	23 - 87	21 - 91	33 - 69	34 - 74	39 - 71
	Oupia	36 - 64	25 - 62	35 - 71	34 - 67	36 - 69
	Saint Simon	28 - 83	28 - 95	23 - 97	34 - 80	41 - 72

Regarding the mean NPRI for class 1 - 5, the largest values are found for day 4 and the smallest for day 1 for all three farms. This is thus in accordance with the trend found in Figures 3.7 and B.3.

The values for day 1 are rather high for Mardyck and Saint Simon while the largest ratios are found for day 3 for Oupia. In the case of Mardyck, the ratio for day 1 is 50 % higher than the ratio for day 2 and it decreases then with look-ahead time. The results for day 2 and 3 are rather similar in the Oupia and Saint Simon case.

The IQRs show the same behaviour with similar results for the first two days followed by narrower distributions further on in time. These narrower distributions come primarily from smaller maximum IQR but also to some extent from larger minimum IQR.

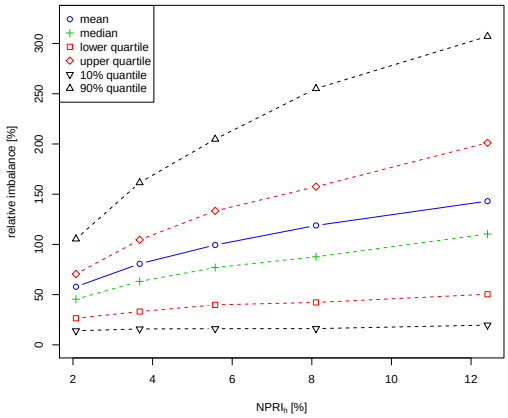
The results show that the ability of NPRI to distinguish between more and less predictable situations decreases with look-ahead time. The usefulness of the index is therefore largest for the first 2 - 3 days. This is however enough, since it is only the first 2 - 3 days that are of consideration for the purposes of risk indices in this work.

3.2.3 NPRI using different lengths of the look-ahead time window

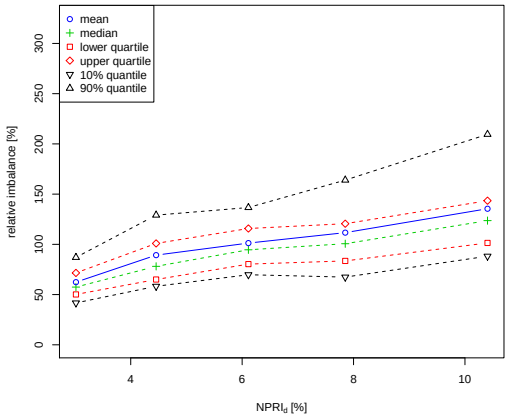
Beside moving the look-ahead time window, the length of the window also requires attention. In order to get enough observations for the halved window, the NPRI is calculated on predictions made with a 1 hour resolution. This is also carried out when the window is doubled to compare the results.

Longer look-ahead time window

When using a longer look-ahead time window, the distribution of relative energy imbalance for each class gets narrower, as observed in Figure 3.9 where $NPRI_h$ and $NPRI_d$ are calculated for day 2 to 3, from look-ahead times between 24 and 72 hours for Saint Simon.



(a) $NPRI_h$



(b) $NPRI_d$

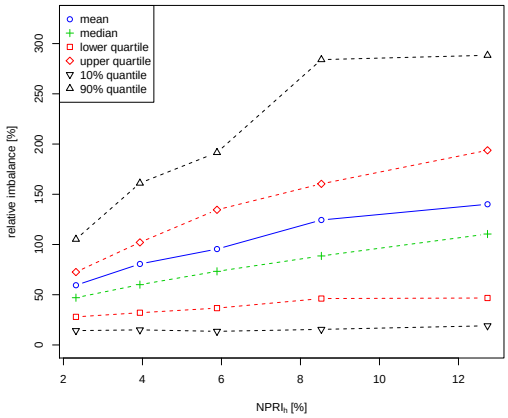
Figure 3.9: $NPRI_h$ and $NPRI_d$ for the 48 hours covering day 2 and 3 for Saint Simon.

The ratio of the mean imbalance level is reduced, especially for $NPRI_d$ where it is now equal to 2.17. The IQRs are also reduced, even here significantly more for $NPRI_d$ where they are reduced to 21 - 42 % giving a smaller difference in sharpness between the classes.

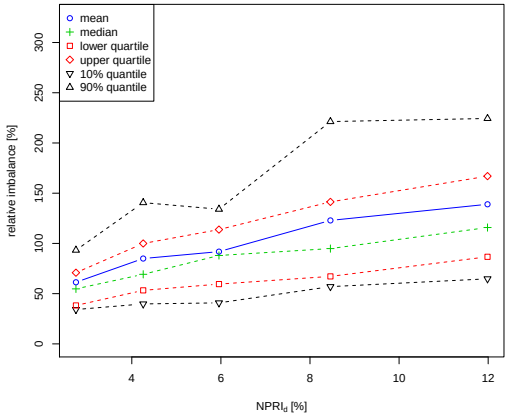
Shorter look-ahead time window

Using a shorter look-ahead time window leads to wider distributions which can be seen in Figure 3.10 where the $NPRI_h$ and $NPRI_d$ are calculated for the first 12 hours of day 3 for

Saint Simon.



(a) $NPRI_h$



(b) $NPRI_d$

Figure 3.10: $NPRI_h$ and $NPRI_d$ for the first 12 hours of day 3 for Saint Simon.

The behaviour for the other two farms is similar and statistics of the results are shown in Table B.1 in Appendix B. The mean NPRI values for class 1 and 5 increases with reduced length of the look-ahead time window. The ratio of mean imbalance level also increases with shorter time resolutions for both Mardyck and Oupia while varying results are found for Saint Simon.

To summarize, one can conclude that using longer look-ahead time windows results in narrower distributions, especially for $NPRI_d$. This implies that the index has a higher skill in predicting the imbalance level on a longer look-ahead time windows which can simply be explained by the fact that energy imbalances are evened out.

3.3 Extension of the NPRI to larger spatial scales

The extension of risk index to a region can be made by calculating the total or average index value for a number of wind farms. Another option proposed here is to produce risk index maps, where a risk index is displayed for each NWP grid point on a map.

3.3.1 Risk indices for the aggregation of wind farms

The average error is reduced when forecasting the aggregated power for several wind farms. There can however still be large errors, for example if there are shortage or surplus at all the farms at the same time. It is therefore of interest to investigate whether the NPRI can be useful in providing information about the level of uncertainty about the aggregated power for several wind farms.

This is investigated by considering the three French wind farms as an aggregate or region. An aggregate of three farms situated so far away from each other is not a realistic case for the integration of the electricity into the grid by TSOs. It is however realistic for an owner to have three wind farms spread out over such a large area. Approximation of the farms into a region is therefore a possible scenario. The easiest way to obtain regional forecasts is to consider the three farms as a single wind farm by aggregating the predicted and measured power for all farms. Details in the reduction in forecast error when considering the three farms as a regions are shown in Appendix B.

Calculating NPRI for the aggregated power of several wind farms is here proposed to be performed in two ways, either by calculating the NPRI on the aggregated forecasts of the farms or by using a weighted mean of the individual NPRI values.

NPRI for aggregated power forecasts

The first of these options is considered by calculating the NPRI on the aggregate of the ensemble predictions. The ensemble members are added member wise according to their position in the array. The energy imbalances are then calculated using the sum of the control members and the sum of the observed power. The obtained NPRI plots for $NPRI_h$ and $NPRI_d$ for day 2 are shown in Figure 3.11.

The ratio of mean imbalance level is reduced to about 1.7 for both $NPRI_d$ and $NPRI_h$. This means that the ability of risk index to distinguish between high and low level of errors is reduced compared to the individual farms. It is however important to remember that the errors, and thereby the energy imbalance levels, also are reduced when the farms are aggregated together. The smaller IQR ranges observed are probably also due to the smaller imbalance levels. The trend of a quasi-linear increase in relative imbalance level with class number is present even though there is not always an increase in the IQR range for $NPRI_h$ when going up one class. Notable is also that the NPRI values are slightly reduced, indicating that the spread of the ensemble members is lower when they are aggregated.

An important remark is that the ensemble members have been added member-wise, implying that it is not certain that the largest and smallest members have been added respectively. Weighting the NPRI from different farms can therefore be assumed to be a better solution.

Weighting of individual NPRI for the aggregate of several wind farms

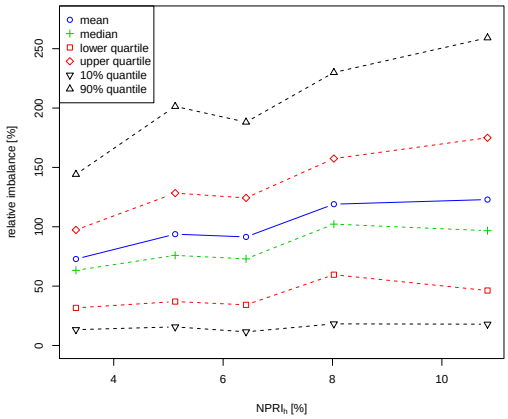
In order to get a representative weighted mean of the NPRI for the three farms, the proposition is to determine the weights by the average contribution of each individual wind farm to the total production of the aggregate of the farms. The contribution of the wind farm m to the total NPRI value is therefore determined by the average production $\overline{P}^m$ of the farm divided by the nominal power P_n^m of the farm giving the following expression for the total NPRI:

$$NPRI = \frac{\overline{P}^m}{P_n^m} NPRI^m \frac{1}{\sum_{i=1}^{M} \frac{\overline{P}^i}{P_n^i}} \tag{3.2}$$

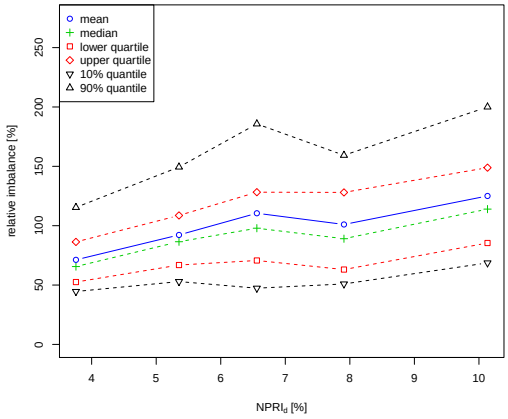
where M is the total number of wind farms, in this case 3.

Results of the NPRI on the aggregate of several wind farms

The total energy imbalance for the three farms is calculated as the sum of the individual energy imbalances. The imbalances are then normalized by the average total energy imbalance for



(a) $NPRI_h$



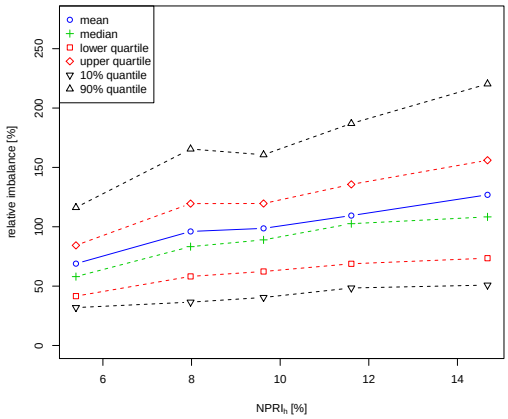
(b) $NPRI_d$

Figure 3.11: $NPRI_h$ and $NPRI_d$ for day 3 when the power is aggregated for the three farms.

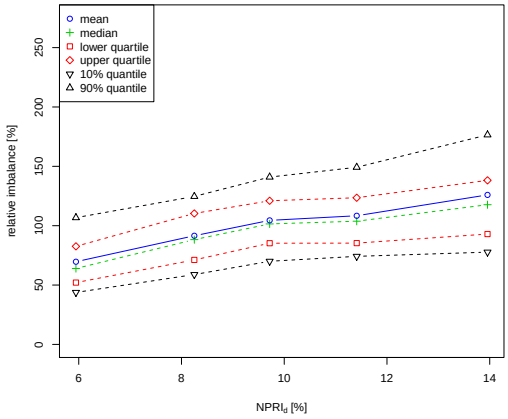
the farms in order to display relative imbalances. Figure 3.12 presents $NPRI_h$ and $NPRI_d$ for day 3.

The ratio of mean imbalance level is slightly improved compared to when the NPRIs were calculated on the aggregated forecasts. The distributions are narrower, especially for $NPRI_d$, but the IQR does not always increase with class for $NPRI_h$. The IQRs are rather similar for all classes showing that there is no large difference in the uncertainty about the energy imbalance depending on NPRI value.

The aggregation of wind farms reduces the forecast error and thereby the energy imbalances. However, the spread of the ensembles is also evened out which has for effect that the NPRI does not give as much information about the uncertainty of the situation. In the considered case, one of the wind farms, Oupia, is situated far away from the other two. The weather can therefore differ significantly between the farms; there can for example be a high pressure front giving low prediction uncertainty for one farm and a low pressure front giving high uncertainty for the other two farms. The total uncertainty can therefore be evened out. Furthermore, a situation where the NPRI is large for all three farms does not necessarily mean that the error will be large since positive and negative error cancel out to some extent.



(a) $NPRI_h$



(b) $NPRI_d$

Figure 3.12: Weighted average of $NPRI_h$ and $NPRI_d$ with total relative energy imbalance for the three farms.

3.3.2 Methods to produce risk index maps

Risk index maps can be used to give information about the level of uncertainty on a regional scale. This can then be used by end-users possessing several wind farms spread out over a larger area to obtain information about where there are higher risks for larger energy imbalances.

Risk index maps are most easily produced using the spread of the wind speed ensembles and not the power ensembles since they are to be produced for every grid point where wind speeds are predicted, independently of whether there is a wind farm there or not. It is therefore of interest to study whether the NPRI based on the ensembles of wind speed can give useful information on the expected energy imbalance.

Results when using wind speed ensembles to compute the NPRI

In order to normalize the wind speeds when calculating the NPRI, the maximum predicted wind speed over all prediction times is used. The results of using the NPRI from the wind speed ensemble forecasts are shown in Table 3.5, where $NPRI_d$ for day 3 and $NPRI_h$ for the last look-ahead time in day 3 are shown for NPRIs based on either ensembles of wind power or ensembles of wind speed. The energy imbalances are in both cases been based on the wind

power forecast control member produced by the RF model.

Table 3.5: NPRI results when using either wind power or wind speed.

	Wind farm	$NPRI_h$ [%] **power**	$NPRI_d$ [%] **power**	$NPRI_h$ [%] **wind speed**	$NPRI_d$ [%] **wind speed**
mean NPRI class 1-5	Mardyck	3.4 - 15.8	4.2 - 14.7	3.4 - 9.5	3.8 - 9.0
	Oupia	5.4 - 18.6	6.6 - 17.2	3.8 - 11.7	4.4 - 11.1
	Saint Simon	2.7 - 13.6	3.4 - 12.3	3.4 - 9.4	3.9 - 8.8
ratio of mean imbalance level	Mardyck	2.45	2.08	1.75	1.76
	Oupia	1.92	2.01	1.85	1.76
	Saint Simon	2.67	2.50	1.91	2.01
minimum and maximum IQR	Mardyck	43 - 160	33 - 69	59 - 132	40 - 69
	Oupia	58 - 136	35 - 71	59 - 129	42 - 70
	Saint Simon	41 - 167	23 - 97	58 - 158	34 - 101

The table shows a decreased resolution when calculating the NPRI on the wind speed compared to the power. This is shown by lower NPRI values, indicating lower spread, and a smaller ratio in mean imbalance level in class 5 and class 1. As an example, calculating $NPRI_d$ for Saint Simon gives a ratio of mean imbalance level in class 5 and 1 equal to 2.05 for the power NPRI and equal to 2.01 for the wind speed NPRI. Similar differences are obtained for the other two farms. The ranges of IQR are also increased for the NPRI calculated from wind speeds, indicating that it has lower skill in predicting the expected level of energy imbalance when compared to the NPRI computed on power ensembles.

The results show that the NPRI calculated from wind speed ensembles is less useful in distinguishing between situations with low and high energy imbalance than the NPRI calculated from power ensembles. This can partly be explained by the fact that the wind direction is also used as an explanatory variable when producing power ensembles.

Using the maximum predicted wind speed as a normalization factor perturbs the comparison to some extent since it is not directly comparable to the nominal power used when calculating the power NPRI. To get a fairer comparison, the maximum predicted power output can be used instead of the nominal power as normalization factor. This however leads to no major differences and the maximum wind speed and nominal power can therefore be used as respective normalization factors when comparing NPRI from wind speed ensemble and power ensembles.

Correlation between spread of power and spread of wind speed

Even though the spread of the wind speed gives worse results than the spread of the power when determining the energy imbalance, it is still of importance to further investigate the usefulness of the ensembles of wind speed. One way to investigate this is to study whether the spread of the wind speeds is correlated to the spread of the power. Figure 3.13 shows the correlations between energy imbalance and $NPRI_d$ for power and wind speed. The $NPRI_d$'s are calculated for day 3 for Saint Simon.

The correlation between the two $NPRI_d$ is rather strong, equal to 0.83. The one calculated on the wind speed ensembles has however a lower correlation with the energy imbalance, as was also found in the NPRI plots. The high correlation between the two $NPRI_d$ is important since it shows that when the spread of the wind speeds is large, so is the spread of the power. Even though the spread of the wind speeds is less useful in informing on the expected relative imbalance, they can still be useful in informing on relative differences in spread between different parts of a geographical area.

Such information can be displayed on risk index maps. Maps are an efficient way of providing information of both the forecast value and associated uncertainty for larger areas [17].

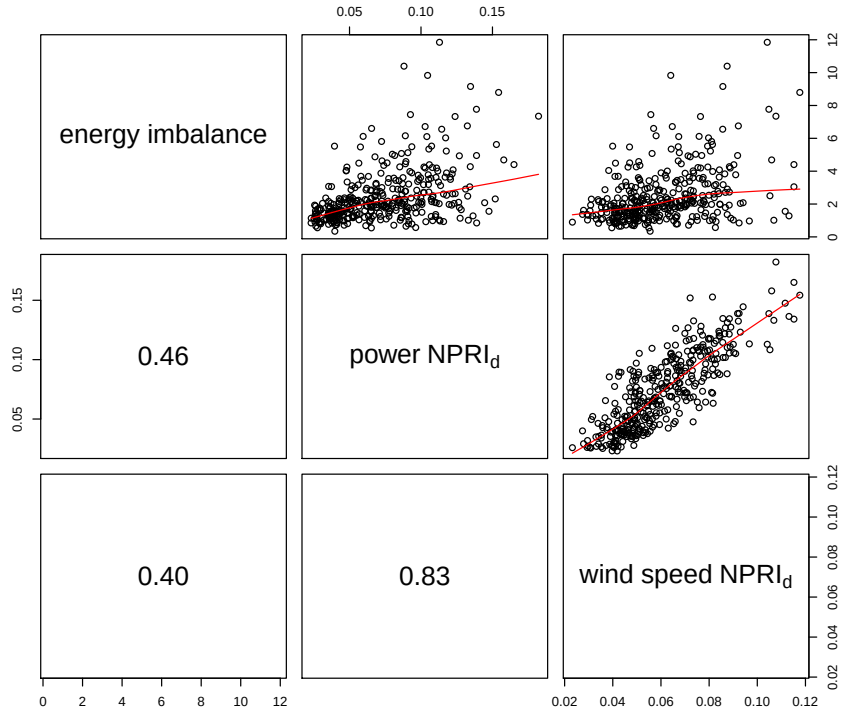


Figure 3.13: Correlations between energy imbalance, power $NPRI_d$ and wind speed $NPRI_d$ for day 3 for Saint Simon.

When providing risk index maps, colour-scales of risk indices can then be used to illustrate the uncertainty in different regions. The wind speed must however also be given since the level of the wind speed has a high implication on the spread when the wind speeds are converted to power.

3.4 Conclusions

This chapter has focused on evaluating the NPRI in the same manner as in [22] as well as extending its examination to other spatial and temporal scales. Compared to previous work in the area, it was found that the NPRI based on the RF model predictions gives a lower distinction between high and low energy imbalances than the NPRI based on the PC model. These results show that in order to extract more information from the ensemble spread it is important to use a prediction model that preserves the spread of the meteorological ensembles when they are converted to power. Using the PC model, based on the theoretical power curves of the farms, is a method that gives larger distinctions between situations with high and low energy imbalances. A large drawback with the PC model is however that the prediction errors are larger.

An alternative approach was therefore proposed consisting of using the NPRI calculated on the power ensembles obtained by the PC model to provide information on the energy imbalances from the control member of the RF model predictions. This was shown to give similar results as using only the RF model. The conclusion of this is therefore that the RF model should be employed since it gives lower average forecast errors. Further investigation of such combinations of models is however interesting to make, for example using a more advanced model than the PC model but with the same spread preserving properties.

Concerning temporal ensembles and multi temporal ensembles it was shown that these were

less useful than standard power ensembles in giving information about the energy imbalance level. Temporal ensembles can however still be a good option, especially since they do not require NWP ensembles, only single NWP. From the choice of weights it was found that using exponentially decaying weights with β equal to 12 or 24 gives best results in terms of highest resolution and sharpness from the NPRI. Since weights with $\beta = 12$ were found to be most suitable in terms of NRMSE and NMAE, it can be concluded that a weight distribution with β is a good alternative when using temporal ensembles.

Furthermore, it was investigated how the ensemble spread evolves in time. It was confirmed that the RF model reduces the ensemble spread when the wind speed ensembles are converted to power ensembles for longer look-ahead times. The NPRI was found to be most useful for the first 2 - 3 days ahead. Concerning the length of the look-ahead time window, it can be concluded that the previously used length of 24 hours, employed preferably on the first 3 days is a good option.

The extension of risk indices on larger spatial scales was proposed using three options. The method to weight NPRI values from several wind farms into a total NPRI value and link this to the forecast for the aggregated power for the farms was proposed. It was found to be better than basing the NPRI on the aggregate of the ensemble members for the farms. Further investigation is however motivated here, for example by aggregating the ensemble members after sorting them according to their value.

Even though the risk index have been applied using predictions made with different models and extended to other temporal and spatial scales, only the NPRI has been evaluated. One of the main purposes of this work is to investigate alternative risk indices and this is therefore performed in the next chapter. The next chapter also contains the investigation of another important purpose of the work, namely how risk indices can be presented to end-users.

Chapter 4

Definitions of Alternative Risk Indices and Presentation of Risk Indices to End-users

The existing definition of risk index, NPRI, might not be the most appropriate way to extract the information provided by wind power ensembles. A couple of proposals for alternative risk indices, among which two are modifications of the existing NPRI, are therefore presented in this chapter. These indices are then evaluated in the same way as the NPRI. They are also investigated jointly since it is relevant to examine whether a combination of two or several indices can be useful.

There exist several actors interested in wind power forecasts such as TSOs, energy service providers, wind farm operators and energy traders. Depending on the requirements from end-users, the information from risk indices needs to be provided in different ways. While some end-users want the total distribution of risk indices others might just need a scale consisting of risks for a larger energy imbalance than usual. It is therefore important to investigate how the information obtained from risk indices can be presented by examining the division into classes and how statistics of the distribution of energy imbalances can be used to inform on risks for large energy imbalances. Alternatives to the previously used way of presenting risk indices are therefore proposed and evaluated in this chapter. An investigation of how risk indices can be used to take decisions is also performed.

4.1 Proposals of alternative risk indices

Before proposing risk indices that differ completely from the NPRI, two different modifications of the NPRI are suggested. The different alternative indices are presented in two parts; indices that are based on each look-ahead time separately and indices that are based on the evolution of ensemble members in time.

4.1.1 Modifications of the existing NPRI

Two modifications of the existing NPRI are considered: changing the weighting of the ensemble members when calculating the weighted standard deviation, Equation 1.1, and using the control member instead of the ensemble mean as the spread reference.

Giving more weight to the control member

In Section 2.4 in Chapter 2 it was confirmed that the control member was in average the most accurate forecast among the ensemble members for ensemble predictions made by the RF model. A reasonable alternative to the equal weighting of ensemble members is therefore

to give more weight to the control member when calculating the weighted standard deviation, Equation 1.1, which is the base of the NPRI.

The weights are constructed by determining a weighting factor, stating how much more than the rest of the ensemble members the control member shall be weighted. In this case, with 51 ensemble members, a weighting factor of 5 gives the weights 1/55 to the first 50 ensemble members and the weight 5/55 (=1/11) to the control member. It is then interesting to experiment with the weighting factor in order to determine whether some weighting might lead to better results than the equal weighting used until now.

Using the control member instead of the ensemble mean when computing the NPRI

So far, the weighted squared deviations between all ensemble members and the ensemble mean have been calculated in the NPRI. The ensemble mean is however always situated in the "middle" of the ensemble set as observed for example in the spaghetti plots in Figure 2.11 in Chapter 2. Contrary to this, the control member can lie close or far from the middle of the envelope of ensemble members. Calculating the weighted squared distances between each ensemble member and the control member is therefore a reasonable modification of the NPRI. This option can be assumed to better distinguish between situations with low and high spread among the ensemble members and skewed distributions of members. The equation for the weighted standard deviation then takes the form:

$$\widetilde{\sigma}_{t,k} = \left[\frac{J}{J-1} \sum_{j=1}^{J} w_j \left(\widehat{P}^{(j)}_{t+k|t} - \widehat{P}^{*}_{t+k|t} \right)^2 \right]^{1/2} \tag{4.1}$$

where $\widehat{P}^{*}_{t+k|t}$ is the control member of the J ensemble members.

4.1.2 Risk indices based on single look-ahead times

For the NPRI, the index is evaluated on each look-ahead time individually and an average over a look-ahead time window is then taken in the case of $NPRI_d$. The first alternative risk indices presented here are similar in the sense that they are also calculated on each look-ahead time separately.

The Max-min index

The *Max-min* index is a simple index that uses the difference between the maximum and minimum values of the J ensemble members, including the control member, for each prediction time t and each look-ahead time k as the measure of the spread of the ensemble members. The difference is normalized by the nominal power P_n, in order to get a scaled index, giving index values between 0 and 1. The index can then be used in a similar way as the NPRI by calculating the mean of the differences between maximum and minimum of the ensemble members over a look-ahead period $[k_1, k_2]$:

$$Maxmin(t, k_1, k_2) = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \frac{1}{P_n} \left(\max_j \widehat{P}^{(j)}_{t+k|t} - \min_j \widehat{P}^{(j)}_{t+k|t} \right) \tag{4.2}$$

Note that when $k_1 = k_2$, the index is calculated over a single look-ahead time k, in the same way as $NPRI_h$ and when $k_2 - k_1 = 24$, assuming a temporal resolution of 1 hour, the index is comparable to $NPRI_d$.

Willmott's Index of Agreement for specific look-ahead times

In the case of NPRI, the level of power output is not considered when calculating the weighted standard deviation. It could however be preferable to use an index that takes this into account in some way.

Willmott's Index of Agreement (IoA), as proposed by Willmott [25], is an index to assess the agreement between series of predictions and observations. It gives supplementary information than just the difference between the predictions $\widehat{P}$ and the observations P since it includes the mean of the observations $\overline{P}$. It takes the form [26]:

$$d_\gamma = 1 - \frac{\sum_{i=1}^{N} \left(\widehat{P}_i - P_i\right)^\gamma}{\sum_{i=1}^{N} (|\widehat{P}'_i| + |P'_i|)^\gamma}, \gamma > 0, 0 \leq d \leq 1 \tag{4.3}$$

where $\widehat{P}'_i = \widehat{P}_i - \overline{P}$ and $P'_i = P_i - \overline{P}$. In [25] γ is set to 2 but a γ equal to 1 can also be used. The index is a descriptive measure and is both relative and bounded, between 0 and 1. A high index of agreement means that the observations and predictions are highly correlated and vice versa.

A proposal in this context is to calculate the index as the average agreement between each ensemble member and all other ensemble members for a specific look-ahead time k. It is named *Willmott for horizons* and takes the form:

$$d_{t,k} = \frac{1}{J} \sum_{i=1}^{J} \left[1 - \frac{\sum_{j=1}^{J} \left(\widehat{P}^{(j)}_{t+k|t} - \widehat{P}^{(i)}_{t+k|t}\right)^\gamma}{\sum_{j=1}^{J} (|\widehat{P}^{(j)}_{t+k|t} - \overline{P}^{J}_{t+k|t}| + |\widehat{P}^{(i)}_{t+k|t} - \overline{P}^{J}_{t+k|t}|)^\gamma} \right], 0 \leq d \leq 1 \tag{4.4}$$

where $\overline{P}^{J}_{t+k|t}$ is the mean of the J ensemble members for look-ahead time k and prediction time t, as defined in Equation 1.2. The coefficient γ is set to 1 or 2, a γ equal to 2 is assumed to give more weight to ensemble members which deviate more from each other.

This application of Willmott's IoA is intended to be a direct alternative to the weighted standard deviation used for the NPRI. As with the NPRI, the index could be calculated for a single look-ahead time or averaged over periods of look-ahead times:

$$d_{t,k_1,k_2} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} d_{t,k} \tag{4.5}$$

One main difference between this proposal and the NPRI is that this index takes values between 0 and 1, meaning that it could be more directly interpreted than the NPRI.

4.1.3 Risk indices based on the evolution of ensemble members in time

A situation where the ensemble members are less correlated could be assumed to correspond to a more chaotic situation that tends to be more unpredictable than a situation where the correlation between the members is high. The variance or spread, which is the standard deviation of the ensemble members, can be the same in both cases but if the correlation of the ensemble members is low, one would expect another type of uncertainty then when it is high. Extraction of the correlation between the ensemble members over time could therefore be useful, possibly in combination with a measure of the spread.

Pearson's Correlation Coefficient

One of the best known measures of correlation is Pearson's Correlation Coefficient which gives the correlation between two series X and Y:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right) \tag{4.6}$$

where $\bar{X}$ and σ_X are the mean and standard deviation of X and Y respectively. r take values between -1 and 1 with 1 when X and Y are absolutely positively correlated and -1 in the case

when they are absolutely negatively correlated. A correlation coefficient equal to 0 means that the two series are uncorrelated.

When adapting this to ensemble forecasts, the proposal is to calculate the average pair wise correlation between all ensemble members over a defined look-ahead time window. The correlation between two ensemble members for look-ahead times k_1 to k_2 takes the form:

$$r_{i,j,k_1,k_2} = \left[\frac{1}{k_2 - k_1} \sum_{k=1}^{k_2-k_1+1} \left(\frac{\widehat{P}^{(j)}_{t+k|t} - \bar{P}^{(j)}}{\sigma^j_P} \right) \left(\frac{\widehat{P}^{(i)}_{t+k|t} - \bar{P}^{(i)}}{\sigma^i_P} \right) \right] \tag{4.7}$$

where $\bar{P}^{(j)}$ and σ_P are the mean and standard deviation of the predictions for the look-ahead time period $[k_1, k_2]$ respectively.

For J ensemble members, J^2 correlations are obtained of which the average is then calculated as:

$$r_{k_1,k_2} = \frac{1}{J^2} \sum_{i=1}^{J} \sum_{j=1}^{J} r_{i,j,k_1,k_2} \tag{4.8}$$

This equation serves as the definition for the index, which is named *Pearson* index.

A disadvantage with this method is that it does not take into account the standard deviation of the ensemble members, only the standard deviation between pairs of ensemble members over time. A set of ensemble members can be perfectly correlated but still present a large spread between the members. It is also possible to find a low spread associated to low correlation. It can therefore be necessary to also take into account the standard deviation in some way, for example by combining the index with the NPRI.

Willmott's IoA as a measure of the correlation of ensemble members in time

Willmott's IoA can also be used to evaluate the agreement between ensemble members over a look-ahead time window, in the same way as Pearson's Correlation Coefficient. The IoA between two ensemble members over a look-ahead time window $[k_1, k_2]$ then takes the following form:

$$d_{i,j,t,k_1,k_2} = \left[1 - \frac{\sum_{k=1}^{k_2-k_1+1} \left(\widehat{P}^{(j)}_{t+k|t} - \widehat{P}^{(i)}_{t+k|t} \right)^{\gamma}}{\sum_{k=1}^{k_2-k_1+1} \left(|\widehat{P}^{(j)}_{t+k|t} - \overline{P}^{(j)}_{t+k|t}| + |\widehat{P}^{(i)}_{t+k|t} - \overline{P}^{(i)}_{t+k|t}| \right)^{\gamma}} \right], 0 \leq d \leq 1 \tag{4.9}$$

where $\overline{P}^{(i)}_{t+k|t}$ is the average of the ensemble member i for the look-ahead time window $[k_1, k_2]$ and γ is set to 1 or 2 as previously.

This index is then extended to all ensemble members by averaging over the J^2 alternative combinations:

$$d_{t,k_1,k_2} = \frac{1}{J^2} \sum_{i=1}^{J} \sum_{j=1}^{J} d_{i,j,t,k_1,k_2} \tag{4.10}$$

giving a risk index named *Willmot for periods*.

In the above presented alternative indices, equal weight is given to all ensemble members. Modifications of the indices could however be made by including weighting factors when it is preferred to give different weights to different ensemble members.

The *Max-min-max* index

An adaption of the *Max-min* index to cover a look-ahead time window is to calculate the difference between the maximum and minimum value of the ensemble predictions over a look-ahead time window. This index is defined as the *Max-min-max*:

$$Maxminmax(t, k_1, k_2) = \frac{1}{P_n} \left(\max_{[k_1,k_2]} \left(\max_j \widehat{P}^{(j)}_{t+k|t} \right) - \min_{[k_1,k_2]} \left(\min_j \widehat{P}^{(j)}_{t+k|t} \right) \right) \tag{4.11}$$

This index is expected to be useful in order to inform on the risk for large variations in the power output which can cause large errors.

Chaos indices

As discussed previously, analyzing the peer order changes of the ensemble members over a look-ahead time window is interesting. Cases where the number of changes is large are expected to be more chaotic and represent less predictable situations then when the number of changes is few. A proposed index for evaluating this is called *Chaos index*. This index measures the average number of times the ranks of the ensemble members change in the considered look-ahead window. The index is calculated by first sorting the ensemble members according to their value for each look-ahead time. Each ensemble member is given a rank number, corresponding to its rank among all members. Between two consecutive look-ahead times it is then examined whether the ensemble members have changed rank or not. The number of ensemble members that have changed rank is stored and this is then carried out between all successive look-ahead times in the window after which the average is used as the index value.

An alternative to the first option, referred to as *Chaos normal*, is to calculate the number of rank changes that are larger than a certain number of positions. This is hereafter referred to as *Chaos large changes* and is expected to better capture the most chaotic situations. A third alternative is to calculate the average number of positions that each ensemble member has changed, named *Chaos sum of changes*.

4.2 Results using the alternative risk indices

The alternative risk indices proposed in the previous section are evaluated primarily by comparing them to energy imbalances in the same way as NPRI. Correlations between all indices and with the level of energy imbalance are also computed and scatter plots are presented. The reason for this is that it is of particular interest to examine whether any alternative index is better than NPRI or can be used together with NPRI to extract more information from the ensembles.

4.2.1 Modifications of the NPRI

The two modifications of the NPRI give more importance to the control member and are evaluated in the same manner as the standard NPRI.

Giving more weight to the control member

The results in terms of $NPRI_d$ when giving more weight to the control member using weighting factors (wf) equal to 2, 5 10 and 20 are shown in Table 4.1 for day 2 for the three farms. The results using the standard equal weighting, $wf = 1$, are included for comparison.

Table 4.1: Results of $NPRI_d$ when giving more weight to the control member using five different weighting factors (WF). The results are shown for day 2 for the three farms.

	Wind farm	$NPRI_d$ $wf = 1$	$NPRI_d$ $wf = 2$	$NPRI_d$ $wf = 5$	$NPRI_d$ $wf = 10$	$NPRI_d$ $wf = 20$
mean NPRI class 1-5	Mardyck	2.7 - 13.3	2.6 - 13.2	2.6 - 13.0	2.6 - 12.6	2.4 - 12.0
	Oupia	5.6 - 14.9	5.6 - 14.8	5.5 - 14.5	5.4 - 14.1	5.1 - 13.5
	Saint Simon	2.5 - 11.2	2.5 - 11.1	2.5 - 10.9	2.4 - 10.6	2.3 - 10.0
ratio of mean imbalance level	Mardyck	2.67	2.67	2.68	2.68	2.70
	Oupia	2.04	2.04	2.03	1.98	1.96
	Saint Simon	2.48	2.48	2.47	2.43	2.43
minimum and maximum IQR	Mardyck	22 - 89	22 - 89	22 - 89	22 - 89	22 - 83
	Oupia	26 - 67	26 - 67	26 - 67	26 - 63	29 - 61
	Saint Simon	26 - 95	26 - 95	26 - 95	26 - 90	25 - 90

Giving the control member 2 to 10 times more weight gives similar result as when all the ensemble members are weighted equally. The results vary but the variations are in general small. For even larger weights the behaviour varies however more, in some cases a weighting factor of 20 gives similar results to those obtained with a wf of 1 while it in other cases it slightly deteriorates the results. Since no positive trend is found in the results, the equal weighting is considered as the most suitable option.

Using the control member instead of the ensemble mean when computing the NPRI

The results in terms of $NPRI_d$ and $NPRI_h$ when using the control member instead of the ensemble mean when computing the indices are shown in Table 4.2 for day 3 for the three farms, and compared to the results obtained using the ensemble mean as the NPRI reference. Equal weighting of the ensemble members are used in both cases. The $NPRI_h$ are given for the last look-ahead time in the look-ahead time window.

Table 4.2: NPRI results using the control member instead of the ensemble mean in the equation for the NPRI.

	Wind farm	$NPRI_h$ [%] **mean**	$NPRI_d$ [%] **mean**	$NPRI_h$ [%] **control**	$NPRI_d$ [%] **control**
mean NPRI class 1-5	Mardyck	3.4 - 15.8	4.2 - 14.7	4.0 - 19.3	5.0 - 17.6
	Oupia	5.4 - 18.6	6.6 - 17.2	6.2 - 23.3	7.6 - 21.3
	Saint Simon	2.7 - 13.6	3.4 - 12.3	3.1 - 16.5	4.0 - 14.9
ratio of mean imbalance level	Mardyck	2.45	2.08	2.26	2.12
	Oupia	1.92	2.01	1.78	1.87
	Saint Simon	2.67	2.50	2.46	2.37
minimum and maximum IQR	Mardyck	43 - 160	33 - 69	47 - 173	36 - 79
	Oupia	58 - 136	35 - 71	58 - 136	43 - 73
	Saint Simon	41 - 167	23 - 97	40 - 166	22 - 95

The average NPRI increases when the control member is used instead of the ensemble mean, for both class 1 and 5. This is expected since the control member can lie among the outermost ensemble members thus giving a larger spread. The ratio of mean imbalance level is in general reduced, except for $NPRI_d$ for Mardyck. Using the control member in the weighted standard deviation thus gives a worse distinction between high and low relative imbalance levels than using the ensemble mean. Finally, the IQR ranges do not change significantly.

The results show that even though using the control member gave a lower ratio of mean imbalance levels, it could still be an alternative to consider. It is therefore of interest to study the relation between the alternative options in order to determine whether a combination could be useful. The correlations are shown along with scatter plots in Figure 4.1 for day 2 for Oupia. The correlation coefficients with the energy imbalance are included for comparison. The $NPRI_h$ is calculated on the first look-ahead time in the considered look-ahead time window.

The $NPRI_d$ calculated on the ensemble mean is found to be most related with the energy imbalances. There are strong relations between the two $NPRI_h$ indices and the two $NPRI_d$ indices with correlations equal to 0.97 and 0.92 respectively. This confirms that the indices calculated on the control member give more or less the same information as the indices calculated on the ensemble mean. The correlations with the energy imbalance are thereby also similar.

The correlations between an NPRI and the energy imbalance are lower for the $NPRI_h$ than for the $NPRI_d$ due to the fact that the energy imbalances are based on the whole look-ahead

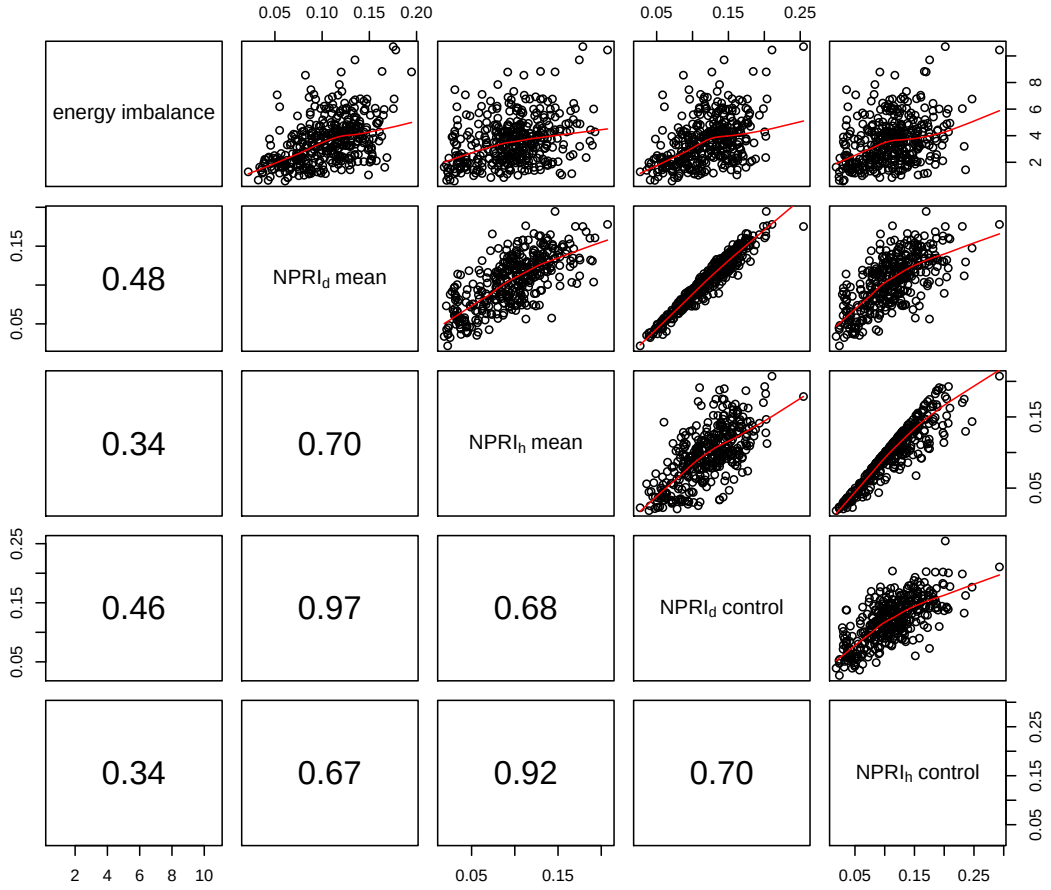


Figure 4.1: Correlation for NPRI based on the mean of ensembles or the control member.

time window and the $NPRI_h$ are only calculated on the first look-ahead time. Comparing the correlations of the different $NPRI_h$ indices in a look-ahead time window with the energy imbalance over the whole window shows no major differences with the correlation found for the first look-ahead time of the window. The correlations vary between look-ahead time windows and wind farms but the deviations are small and no general trend is found. The first $NPRI_h$ in a period can therefore be a good indicator of the expected energy imbalance for the coming period, even though it is not as good as $NPRI_d$.

4.2.2 Alternative risk indices

The alternative risk indices are here evaluated in the same way as with the NPRI previously. Correlations between the indices are also shown.

NPRI results using the *Max-min* and the *Max-min-max* indices

Statistics of the performances of the *Max-min* index and the *Max-min-max* index are shown in Table 4.3 together with the $NPRI_d$ for day 3 for the three farms.

The performances of the indices are found to be relatively good compared to the $NPRI_d$, especially for the *Max-min* index. In terms of ratio of mean imbalance level, the *Max-min*

Table 4.3: Results for $NPRI_d$ and the two max min indices.

	Wind farm	$NPRI_d$ [%]	***Max-min*** [%]	***Max-min-max*** [%]
mean index value class 1-5	Mardyck	4.2 - 14.7	19 - 58	28 - 70
	Oupia	6.6 - 17.2	29 - 65	41 - 78
	Saint Simon	3.4 - 12.3	16 - 49	24 - 63
ratio of mean imbalance level	Mardyck	2.08	2.32	1.93
	Oupia	2.01	1.83	1.64
	Saint Simon	2.50	2.44	2.00
minimum and maximum IQR	Mardyck	33 - 69	36 - 69	43 - 68
	Oupia	35 - 71	40 - 67	48 - 74
	Saint Simon	23 - 97	24 - 95	32 - 87

index performs better than $NPRI_d$ for Mardyck and slightly worse for the other two farms. Considering the sharpness of the distributions, the results are similar for the *Max-min* and the $NPRI_d$. The *Max-min-max* index has narrower distributions for higher classes but also wider distributions for lower classes. The differences between $NPRI_d$ and the *Max-min* index are however generally not large which means that the latter could be used as an alternative to the NPRI.

Since the two indices show similar performance, it is important to know whether they are correlated or not. A strong, linear correlation means that the indices measure roughly the same thing while a non-correlation means that the indices could be complementary.

Correlations between the energy imbalance, $NPRI_d$, *Max-min* and *Max-min-max* indices

Figure 4.2 shows the respective correlations between energy imbalance, $NPRI_d$, *Max-min* and *Max-min-max* for day 2 for Saint Simon.

The *Max-min* index is found to be almost equally correlated with the energy imbalance as the $NPRI_d$ with correlation coefficients of 0.47 and 0.48. Furthermore, the relation between the two indices is shown to be strong. For the correlation between $NPRI_d$ and the *Max-min-max* index, a small trend of a curved relationship is found.

Since the combination of a risk index calculated from the PC model predictions with energy imbalances from the RF model control member was found to give relatively good results, using this approach for the *Max-min* index is worth examining. The results are shown in Figure C.1 in Appendix C. This option gives slightly stronger correlations between the $NPRI_d$ and the energy imbalances and between the *Max-min* index and the energy imbalances compared to when the RF model is used to calculate the indices. This indicates that the indices calculated with the PC model better explains the energy imbalances from the RF model than the indices from the RF model. An almost perfect positive correlation between the $NPRI_d$ and the *Max-min* index is also found.

The correlations confirm thereby that the distance between the maximum and minimum of the ensemble members gives almost the same information as the standard deviation of the ensemble members in informing on the level of energy imbalance. Since the distance between the maximum and minimum value is easier to interpret for an observer than the standard deviation, the *Max-min* index can be a good alternative to use when informing about the expected level of energy imbalance and the relative uncertainty of the predictions.

Results using the Chaos indices

Correlations between the energy imbalance, the $NPRI_d$ and the three versions of the *Chaos index*, with *Chaos large changes* considering changes larger than 10 positions, are shown in Figure 4.3 for day 2 for Mardyck.

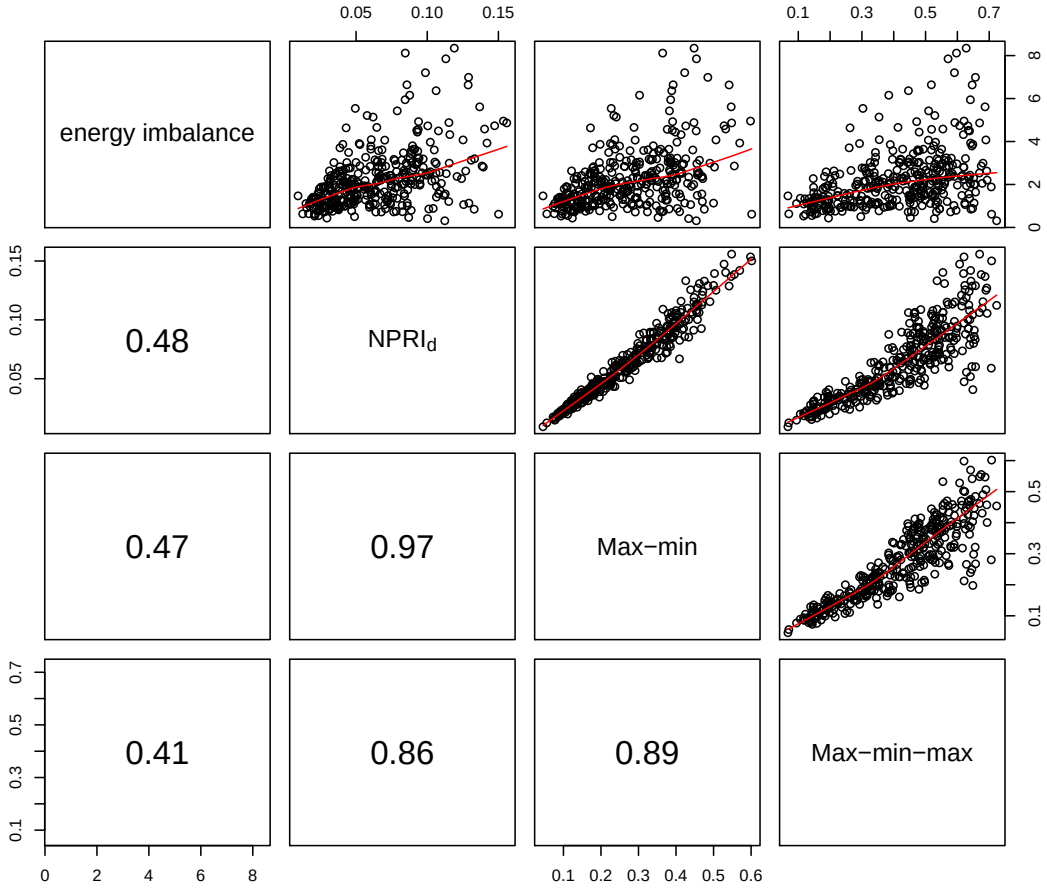


Figure 4.2: Correlation between energy imbalance, $NPRI_d$, the *Max-min* index and the *Max-min-max* index for day 2 for Saint Simon.

None of the alternative chaos indices are shown to give much information on the energy imbalance. For the *Chaos normal* index, the difference in the number of changes for different situations is very small giving a very low correlation with the energy imbalance. *Chaos large changes* and *Chaos sum of changes*, have a negative correlation with the ensemble spread, $NPRI_d$. This means that when there are many large rank changes, or the sum of the rank changes is large, there is also a low NPRI and vice versa. This is contrary to the expected behaviour but can be explained by the fact that it is easier for the ensemble member to cross each other's trajectories when they lie close together then than they are far apart. The two alternatives: *Chaos large changes* and *Chaos sum of changes* are also found to be highly correlated with a correlation coefficient equal to 0.94 indicating that they measure very similar phenomena.

Results using the remaining alternative risk indices

Willmott for horizons is an index that is supposed to be similar to NPRI but weighted, thus being more comparable from case to case. Results, using a γ equal to 2, show however that the index does not give any direct information on the energy imbalance.

Willmott for periods and the *Pearson* index are two indices that measure the average pair-wise correlation between the ensemble members over a specific look-ahead time window. None of the two is found to give any direct information on the energy imbalance over the same

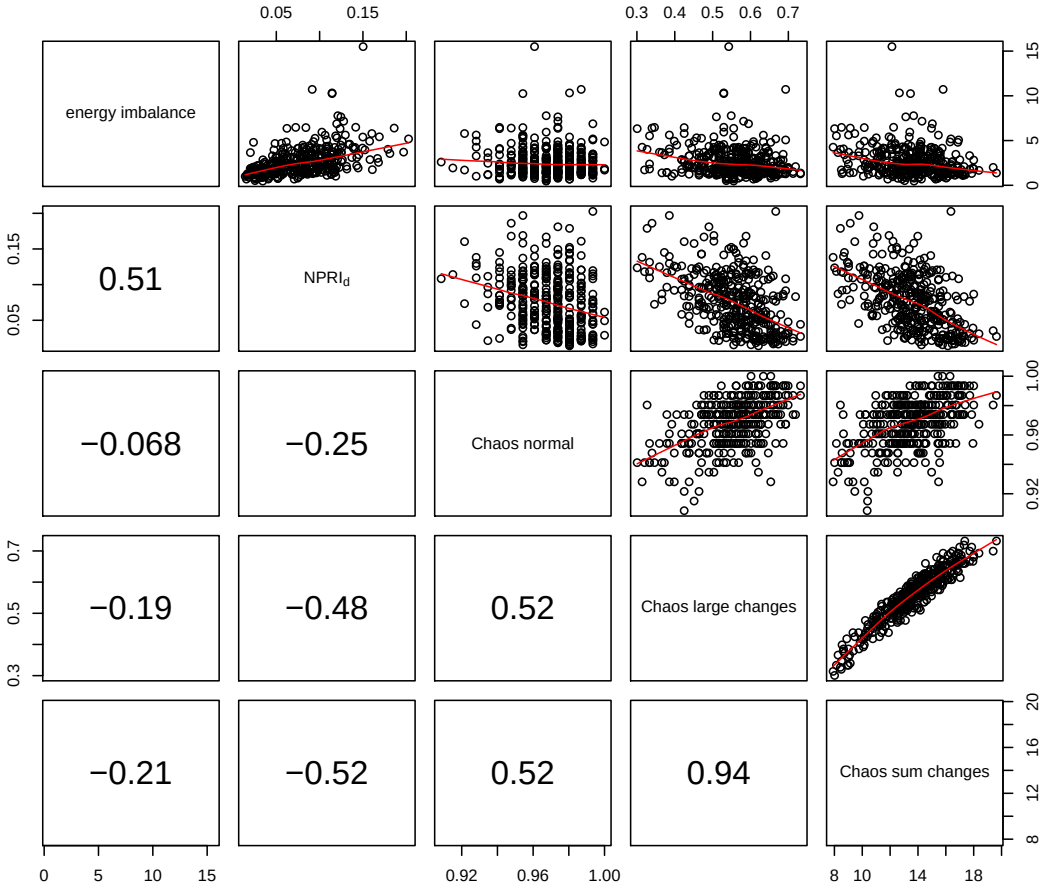


Figure 4.3: Correlations between the $NPRI_d$ and the three variations of the *Chaos index* considering changes larger than 10 positions for the *Chaos large changes* index.

period.

Correlations between *Willmott for horizons*, *Willmott for periods* and the *Pearson* index are shown in Figure 4.4 for Saint Simon for day 3. The parameter γ is set to 2 for the two versions of the Willmott index. Correlations with energy imbalance and $NPRI_d$ are also included. Correlations when the indices are calculated on the predictions made by the PC model are shown in Figure C.2 in Appendix C.

No particular correlations are found between the energy imbalances and any of the alternative risk indices presented in the figure. The strongest correlation with the energy imbalances are found for the $NPRI_d$. This shows that none of the new indices are useful in extracting more information from the ensembles than what the NPRI does. From the scatter plots it is apparent that there is not trend indicating that a combination of any of the indices could be useful. *Willmott for periods* and the *Pearson* index are however found to be highly correlated, showing that they measure roughly the same behaviour of the ensembles. This can however be expected since both indices are measures of the correlation.

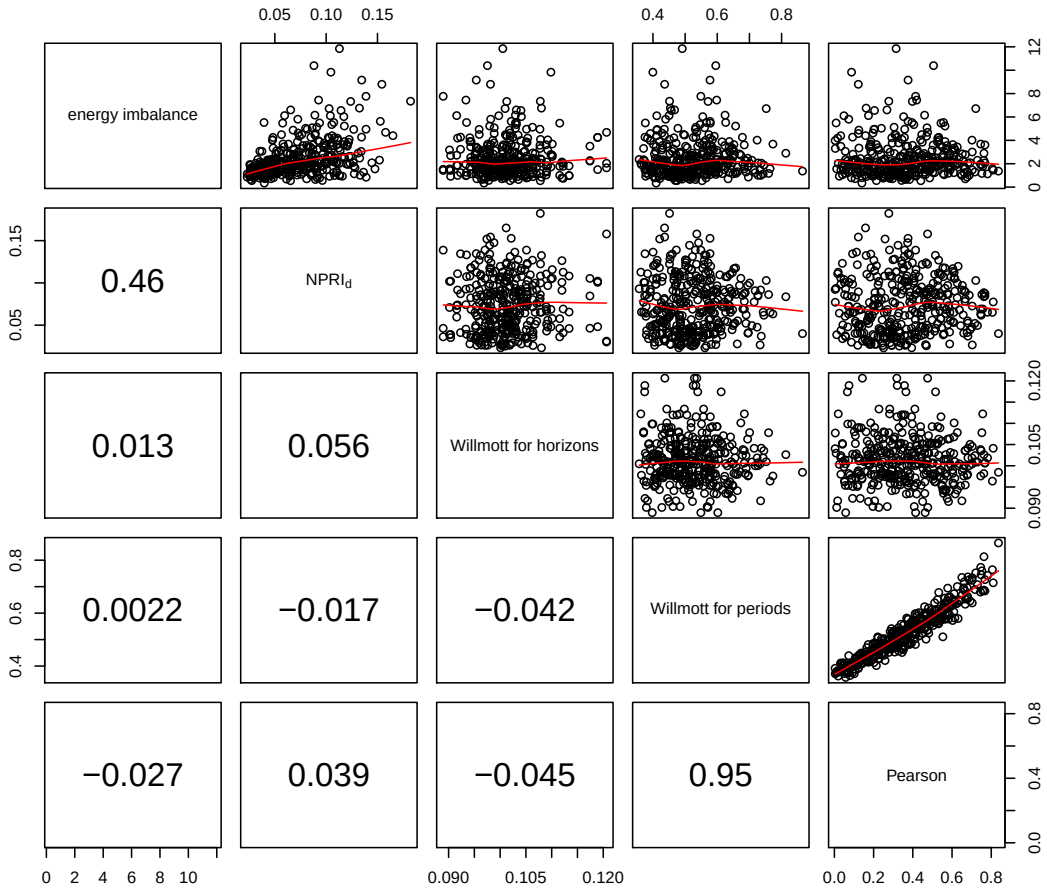


Figure 4.4: Correlations between energy imbalance, $NPRI_d$, *Willmott for horizons* index, *Willmott for periods* index and *Pearson* index for day 3 for Saint Simon.

4.3 Proposals of alternative methods to present risk indices

In [22] and so far in this work, risk indices have been divided into 5 equally populated classes and statistics of the distribution the energy imbalances in each class have been given as measures about the expected level of energy imbalance and uncertainty of the prediction. Examination of how risk indices values are divided into classes and how statistics of the distributions can be obtained motivates the investigation of other alternatives.

4.3.1 How risk indices are divided into classes

The first alternative to consider is the number of classes used to group the risk indices values. This is investigated by increasing the number of classes to 10 and comparing this with the results obtained using 5 classes.

The second alternative to 5 equally populated classes consists of using 5 equally spaced classes. The classes are obtained by dividing the range of risk indices values, the distance between the maximum and minimum value, into 5 equal parts. This method can be motivated by the observation from the scatter plots where energy imbalances are plotted against NPRI values, for example Figure 4.1 and Figure 4.4, that the majority of NPRI values are found in

the low or medium range, resulting in a left-skewed distribution.

4.3.2 Using a sliding window

A drawback with the division of NPRI values into discrete, pre-determined classes is that values close to a boundary of a class might not fit well in either of the two classes. An alternative to avoid this problem is to use a sliding window where risk index (RI) values in $[RI-\Delta_{RI}, RI+\Delta_{RI}]$ are considered for the obtained RI value. Statistics of the distribution of energy imbalances corresponding to the RI values in the window can then be given to inform on the expected energy imbalance. The size of the window, given by $2 \cdot \Delta_{RI}$, needs to be tuned in order to get enough values for an accurate statistical estimation of the energy imbalances while keeping the window relatively small to increase precision.

A way to evaluate the sliding window approach is to divide the set of RI - energy imbalance pairs into two sets, one for learning and one for testing. For each RI value in the testing set, energy imbalances in the learning set whose corresponding RI values lie in the related sliding window are then considered. Statistics of these energy imbalances are calculated and plotted to give information about the expected energy imbalance for the considered RI value. The observed energy imbalances can then also be included in order to validate the results.

4.3.3 Estimation of the quantiles of fitted distributions

So far, only the mean and quantiles of the relative energy imbalances corresponding to specific NPRI values have been given. This gives however only partial information about the distribution of imbalances. A more advanced approach could be to fit distributions, such as Gaussian, beta, or chi-squared, to the energy imbalances for each class. This could be made both on discrete classes and using sliding windows. If a good fit is obtained with a distribution, this would be a more informative alternative than just the quantiles.

Once fitted distributions are obtained, estimations of the upper tails of the distributions, which represents the probability for large energy imbalances can be made. These upper tails can be estimated by statistical methods such as Value at Risk (VaR) and Conditional Value at Risk (CVaR) which are two common risk measures in finance and insurance.

Value at Risk

VaR is associated to the expected loss L of a portfolio and is defined, for a given level of α, as the lowest value of x such that the probability that the loss L exceeds x is $(1-\alpha)$ [7].

$$VaR_\alpha = inf\left\{x \in \mathbf{R} : p(L > x) \leq 1-\alpha\right\} \tag{4.12}$$

In this context VaR_α is proposed to be interpreted as the lowest energy imbalance level, D_{min}, such that the probability that the energy imbalance $D_{t+k_1}^{t+k_2}$ for the look-ahead time window $[k_1, k_2]$ will be larger than D_{min} is $(1-\alpha)$.

$$VaR_\alpha(k_1,k_2) = inf\left\{D_{min} \in \mathbf{R} : p(D_{t+k_1}^{t+k_2} > D_{min}) \leq 1-\alpha\right\} \tag{4.13}$$

Conditional Value at Risk

CVaR is also called expected shortfall and is defined as the conditional expectation of losses above the level x [7].

$$CVaR_\alpha = E\left[x : x > VaR_\alpha\right] \tag{4.14}$$

It thus gives the mean of the eventual losses above the level x for the probability $(1-\alpha)$.

CVaR was included in a model in [7] where a risk-based decision method for optimizing the benefits of trading of wind power was proposed and evaluated. By using alternative strategies depending on the risk for large losses, the method was shown to increase the revenues for wind energy traders. In this context, $CVaR_\alpha$ could be used to give information about the

expected energy imbalance $E\left[D_{t+k_1}^{t+k_2}\right]$ for the look-ahead time window $[k_1, k_2]$, when the energy imbalance $D_{t+k_1}^{t+k_2}$ is larger than D_{min}.

$$CVaR_\alpha(k_1, k_2) = E\left[D_{t+k_1}^{t+k_2} : D_{t+k_1}^{t+k_2} > VaR_\alpha(k_1, k_2)\right] \tag{4.15}$$

4.4 Evaluation results using different presentation techniques

The alternative presentation techniques presented in the previous section are evaluated using NPRI calculated from the ensemble predictions made by the RF model. Due to the limited time available to perform this study, other types of ensembles and other risk indices are not investigated. For the same reason, fitting of distributions and estimations of VaR and CVaR are not examined.

4.4.1 Using a larger number of classes

A division into 10 equally populated classes is investigated when using the RF model. Results of $NPRI_h$ and $NPRI_d$ are presented in Figure 4.5 for day 3 for the Saint Simon wind farm. The $NPRI_h$ and $NPRI_d$ are made up from predictions interpolated to an hourly resolution in order to include more data in each class.

The ratios of mean relative imbalance level in the first and last class equals 2.77 for $NPRI_h$ and 2.29 for $NPRI_d$. This is an increase for $NPRI_h$ and a decrease for $NPRI_d$ compared to using 5 classes. The IQR ranges are slightly decreased to 39 - 155 % and 15 - 78 %. There is especially a very low IQR in the first class for $NPRI_d$ which is advantageous.

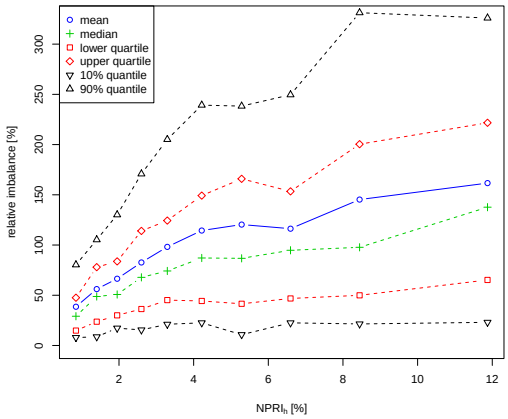
This division into more classes also results in a less linear increase in mean imbalance level between the classes than that found using 5 classes, especially for $NPRI_d$. There are even several cases where the mean imbalance level or some of the quantiles is reduced when augmenting the class number. This probably occurs since the difference between the values in different classes is marginal. This is a drawback with the considered option and indicates that using 5 classes might be preferable. This trend has sometimes been observed even when using 5 classes and diminishing the number of classes is therefore not considered.

4.4.2 Using equally spaced classes

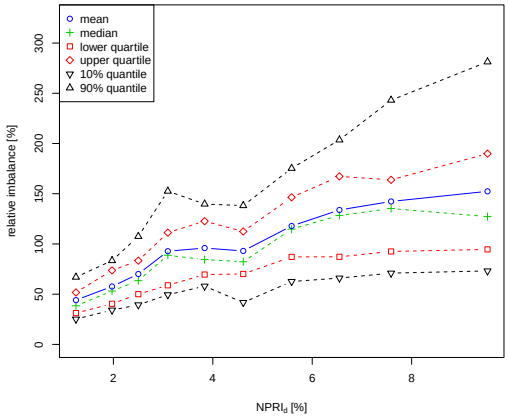
Two examples of $NPRI_d$ using equally spaced classes are shown for Saint Simon for day 2 and day 3 in Figure 4.6. Only $NPRI_d$ is considered since it has been shown in the previous chapter to have a higher skill than $NPRI_h$ in providing information about the relative uncertainty of the forecasts.

Since high NPRI values are less common, the equal spacing results in different numbers of observations in each class. For day 3, presented in Figure 4.6b, 128, 110 and 85 out of the 360 $NPRI_d$ - energy imbalance pairs fall into the first three classes respectively, followed by 31 and 6 observations in the last two classes. The ratio of mean imbalance level is increased significantly, to 3.38 compared to 2.50 using equally populated classes. This shows that this option can give a larger distinction between high and low energy imbalances. Using the terminology presented in Table 1.1 it can for example be observed that it is very unlikely or extremely unlikely to observe an energy imbalance larger than usual when the NPRI value is in the first class. On the other hand, this is very likely or extremely likely to occur when the NPRI is in the last class. For day 2, presented in Figure 4.6a, the number of observations in the last class is higher. In this case, the ratio of mean imbalance level is only slightly increased, to 2.50 compared to 2.46.

Studying the other two farms gives varying results. For Mardyck, the ratio of mean imbalance level is similar to the one found using equally populated classes. For Oupia, the performance is significantly improved with a ratio of mean imbalance level increased from 2.00



(a) $NPRI_h$



(b) $NPRI_d$

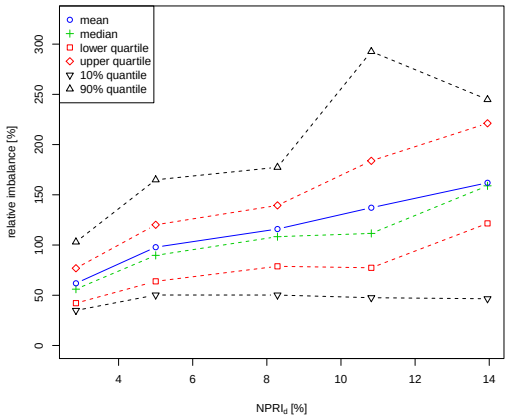
Figure 4.5: $NPRI_h$ and $NPRI_d$ for day 3 for Saint Simon using 10 equally populated NPRI classes.

to 2.81 for day 2 and from 2.03 to 2.45 for day 3. The IQRs are narrower for the first class and wider for the last when compared to equally populated classes. Furthermore it can be remarked that the distribution of NPRI values is more centred in the Oupia case indicating that small NPRI values are less common.

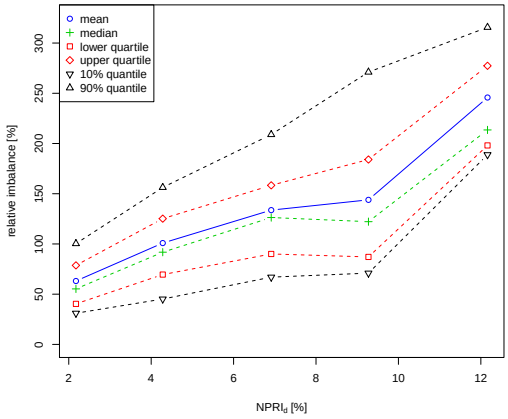
4.4.3 Using a sliding window

Figure 4.7 shows $NPRI_d$ plots for day 2 and day 3 for Saint Simon using a sliding window and dividing the data into learning and testing sets. Three fourths (3/4) of the $NPRI_d$ - energy imbalance pairs are used in the learning set and the approach is evaluated on the remaining quarter, corresponding to 90 observations. In order to obtain quantiles that are increasing, or at least non-decreasing, a relatively large window size is found to be required. Experimenting with window sizes of 1/5 or 1/10 of the range of NPRI values, $[\min(NPRI_d), \max(NPRI_d)]$, shows that a size of 1/5 is preferable. This gives a rather small amount of observations for the largest NPRI values, around 20, while up to 100 observations, more than a third of the available data, are used in the most populated parts. Note that the energy imbalances are normalized by the average energy imbalance computed on the learning set and not on all data.

The first remark to make is that a number of energy imbalances are far higher than the



(a) day 2



(b) day 3

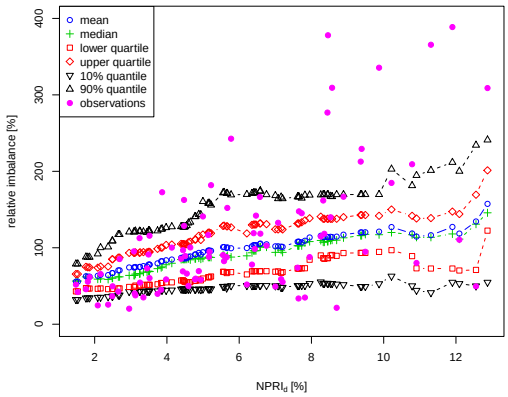
Figure 4.6: $NPRI_d$ day 2 and day 3 for Saint Simon using equally spaced classes.

set of plotted quantiles. There are for example a number of energy imbalances which are 4 - 5 times larger than normal, especially for day 3 in Figure 4.7b. These high imbalances occur mainly for larger NPRI values and indicate that the uncertainty of the expected relative energy imbalance is higher for larger NPRI values.

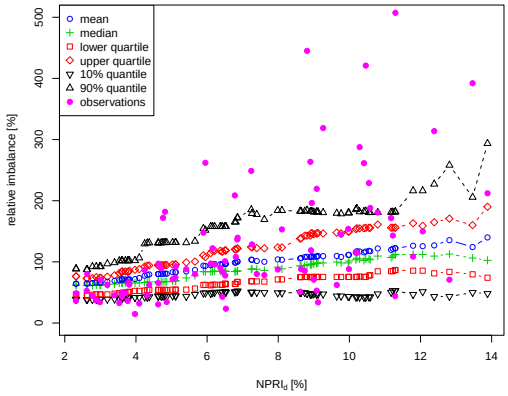
If the statistics were representative for the data, these large values would have been expected in the upper quantiles. The problem is therefore probably due to a lack of data in building up the statistics. Another reason could be that the analysis is made on predictions from the end of May to the middle of December. Using the first 3/4 of this period for the learning set might not be representative for the last quarter since higher wind speeds, giving situations that are more difficult to predict and more dispersed ensemble members, are more common in the autumn than in the summer. Nevertheless, the results indicate that this option can be useful which motivate further studies.

4.5 Using risk indices in decision making processes

Risk indices obtained from ensembles of NWPs can, as discussed in the previous chapters, be an efficient manner to give supplementary information to point predictions. This can enable end-users such as TSOs, wind farm operators or energy traders, to develop alternative strategies



(a) day 2



(b) day 3

Figure 4.7: $NPRI_d$ day 2 and day 3 for Saint Simon using sliding windows and dividing the data into learning and testing sets. Both NPRI and energy imbalances have been computed on the wind power ensembles produced by the RF model.

depending on the uncertainty of the forecast and the risk for large errors.

4.5.1 Confusion tables

One way to give information about the uncertainty in wind power forecast is to use risk indices to give alerts when the probability for a defined energy imbalance is larger than a certain value. As observed in Figure 4.7, it can be recommended to give alerts for large energy imbalances when the $NPRI_d$ is large. To investigate whether the alerts are correct, or whether it was an error to not give an alert, a table with the frequencies of correct alerts and correct false alerts can be made. Such tables, sometimes referred to as *confusion tables* [12], are organized as Table 4.4.

Table 4.4: A confusion table presenting possible errors in statistical decision processes

	Alert needed	**Alert not needed**
Alert made	True alert (TP)	False alert (FP)
Alert not made	False non alert (FN)	True non alert (TN)

The first row of the table presents the two cases that can occur when a decision to make an alert is taken. If this decision is found to be justified, this is referred to as a *true positive* (TP). On the other hand, if this decision turns out to be false, a so called *false positive* (FP) is generated. When a decision to not make an alert is taken, two cases can also occur. If it is found that an alert was needed, a false non alert or *false negative* (FN) occur. If the decision to not make an alert is found to be correct, this is named a *true negative* (TN).

Such tables are often used in statistical decision processes and can in this context be used to compare different visualization options, different risk indices and so forth, by investigating their performance when giving alerts. The option with largest frequencies of correct decisions, TP and TN, would then be preferable. For end-users of a wind power forecast, the different decisions incorporate different strategies. The two types of errors could also be associated to different costs. Whether the chosen method gives few false positives or few false negatives can therefore be considered when determining the best approach.

In order to validate this, a criterion for which a warning should be given must first be determined. An example could be to give alerts when there is a probability superior or equal to y for an energy imbalance $D_{t+k_1}^{t+k_2}$ that is x times larger than usual. Mathematically this can be formulated as:

$$f\left(D_{t+k_1}^{t+k_2}\right) = \begin{cases} 1, & p\left(D_{t+k_1}^{t+k_2} > x \cdot \overline{D}_{k_1}^{k_2}\right) > y \\ \\ 0, & p\left(D_{t+k_1}^{t+k_2} > x \cdot \overline{D}_{k_1}^{k_2}\right) \leq y \end{cases} \tag{4.16}$$

where $\overline{D}_{k_1}^{k_2}$ is the average energy imbalance. The parameters x and y can then be varied depending on the users sensitivity to large energy imbalances. An example could be to give alerts when the probability for an energy imbalance larger than 1.5 times the average imbalance is larger than 20 %.

4.5.2 Evaluation of confusion tables

Using this approach by giving alerts when the probability for an energy imbalance larger than 1.5 times the usual imbalance is larger than 20 % is evaluated on day 2 for Saint Simon. The results are shown in Table 4.6 with the number of the four possible outcomes. A total of 90 cases have been considered.

Table 4.5: Results from giving alerts when the probability for an energy imbalance larger than 1.5 times the normal imbalance is larger than 20 %. The results are given for Saint Simon for day 2.

	Alert needed	**No alert needed**
Alert made	4	1
No alert made	15	70

Alerts were made in 5 of the 90 cases and this decision was found to be correct in 4 out of these 5 cases. When it was decided to not give an alert, this was correct in 70 of the 95 cases corresponding to 73.7 %. The results are thus relatively good. There are particularly a low number of false positives which is advantageous since following the strategy to make an alert can incorporate large costs if this decision turns out to be false.

Using the same strategy for day 3, the results are shown in Table 4.5.

Table 4.6: Results from giving alerts when the probability for an energy imbalance larger than 1.5 times the normal is larger than 20 %. The results are given for Saint Simon for day 3.

	Alert needed	**No alert needed**
Alert made	17	15
No alert made	6	52

The percentage of false positives is larger and the alert is correct in only 15 of the 32 alerts. Observe however that an alert was made in more cases this time. The number of false negatives is still low, representing only 10.3 % of the cases when an alert was not made.

Using this decision option on the other farms gives worse results. For Mardyck, there are more false positives than true positives for both day 2, 13 FP out of 24 alerts, and day 3, 9 FP out of 15 alerts. The number of true negatives is however still high, 82.5 % and 72.1 % of the non alerts. For Oupia, the performance is better with 5 TP out of 8 alerts for day 2 and 16 TP out of 29 alerts for day 3. The number of true negatives is relatively large even here with 68.3 % and 77.0 % of the non alerts.

4.6 Conclusions

In this chapter, a number of alternative risk indices and two modifications of the NPRI have been investigated. It was found that the two proposed modifications of the NPRI and most of the alternative risk indices dit not ameliorate the results. The *Max-min* index was however found to be a good alternative to the $NPRI_d$ showing similar results. Further investigation of this index is therefore motivated, especially since it is easier to calculate and interpret. This should be done by comparing its relative performance to the $NPRI_d$ in several ways and on several test cases.

From the investigation of the division of NPRI - energy imbalance pairs into classes, it can be concluded that equally spaced classes are an interesting option compared to equally populated classes. Further investigation is however needed here, as well as a study of the performance when this option is applied to the alternative risk indices, particularly the *Max-min* index.

The sliding window approach proposed in this work as an alternative to pre-determined classes was found to give good results motivating further investigation. When using this approach, data was also split into a learning set for building up the statistics and a testing set for evaluation. A conclusion of this is that larger data sets covering longer period of time are required in order to obtain learning and testing sets representing similar conditions. The approach of learning and testing sets could equally well be used for pre-determined fixed classes. It is also necessary to investigate the use of a dynamic approach where the statistics in the learning set are updated after every forecast run with the results from last runs since such an approach would be required in online operations.

The evaluation approach using confusion tables presented in the last section of the chapter is an alternative to evaluate the performance of risk indices used to make decisions. It was found to work rather well in these cases but further and more detailed investigation is however required before any conclusions can be drawn. Using larger data sets to build up the statistics with values covering longer time periods are also required for this examination.

Chapter 5

Conclusions and perspectives for future work

This work has been an extension of the work made in [22] that examines in different directions the ideas introduced therein. The work therefore constitutes a broader study in the area of risk indices obtained from wind power forecasts ensembles. Several contributions have been made and many of the ideas proposed and evaluated here merit further investigation.

The main purpose of this work has been to analyze how risk indices, based on wind power forecast ensembles, can be used to give information about the uncertainty in wind power forecasts. The results show that the ensemble spread reflects the uncertainty of the forecast to some extent. Risk indices have been shown to be a manner to capture this spread and link it to the expected level of energy imbalance over single look-ahead times or look-ahead time windows. Such indices are not only able to give an indication of the expected energy imbalance but are also useful to give information about the related uncertainty. Risk indices are especially found to be useful to give an indication about the risk for large forecast errors.

Wind power forecast ensembles are found not to capture the whole range of possible outcomes. The choice of prediction model, particularly how the ensemble spread is preserved when wind power ensembles are derived from NWP ensembles, is therefore of major importance. Further investigation of the use of more advanced prediction models that still preserves the ensemble spread relatively well is necessary. Also of interest is how the spread in the ensembles obtained from WPF models that preserve the ensemble spread are related to the forecast errors from models giving low errors in order to evaluate whether such a combination could be useful. Recalibration of the ensemble members could also be an alternative.

The extension of the previously used definition of a risk index into other spatial and temporal scales show that increasing the scales averages out the forecast errors and makes risk indices less informative. Concerning alternative risk indices, it is found that most of the proposed alternatives do not give better results than the existing definition. Some of them give however interesting results and require more investigation. They way in which risk indices are presented have been investigated briefly showing that alternative methods proposed are promising and merit more attention.

It must be noticed that the NWP ensembles used in this work have a relatively coarse temporal and spatial resolution. The temporal and spatial resolution for single weather forecasts is usually finer, and these are usually used when trying to produce as accurate wind power forecasts as possible. Since the accuracy of NWP is crucial in WPF, improving the temporal and spatial resolutions of NWP ensembles could give more accurate wind power forecast ensembles. This is possible but not certain due to the fact that local models are nested into the boundary conditions of a global model. Errors in the global model can therefore be transferred to the local one. It was found in this work that an interpolation from a 6 hour to a 1 hour resolution works well. It would however be interesting to use predictions made from single NWP with finer resolution and complement them with information about the uncertainty obtained

from ensemble predictions having a coarser resolution. Furthermore, an investigation of the performance obtained using temporal ensembles obtained from a model using NWP with finer resolution is of interest.

The energy imbalances considered in this work are calculated from the absolute errors of the forecasts. No information about the sign of the imbalances is therefore given. For a TSO, shortage or surplus of wind energy causes two different situations, covering the shortage by increasing the production from dispatchable sources or avoiding the surplus by producing less from conventional energy sources. The sign of the error is also important for energy traders, for whom shortages and surpluses can incorporate different costs and penalties when participating on the electricity market. Developing risk indices that take into account the expected sign of the forecast error would therefore be a useful future improvement.

One of the main purposes of risk indices is to inform on the risk for large errors. How to better avoid and warn for large errors is one of the main areas of research in WPF. This can for example be made by warning for *ramp events*, which are situations where the power output presents large variations on a short time interval. It could therefore be interesting to investigate what benefits can be obtained by combining risk indices from ensembles with ramp warning methods.

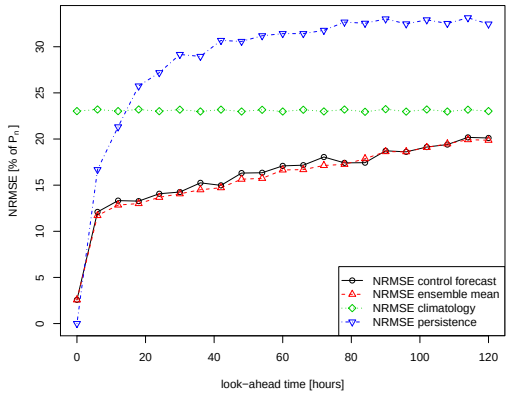
A more global remark in the use of WPF is the usefulness of wind power predictions in general. Even if a prediction model gives low NMAE for a wind farm, if the power factor of the farm is also low, the absolute errors correspond to a larger part of the produced power than for a wind farm with the same NMAE but with higher power factor. The use of the nominal power as a normalization of the power output can therefore be dangerous since it does not show the whole truth. Normalizing by the power factor could be a better idea to give more transparent results. However, even if the average absolute error in percentage of the produced power is high, advanced forecasting models still give significantly better result than climatology or persistence. Continued research in the area of WPF is therefore necessary.

Appendices

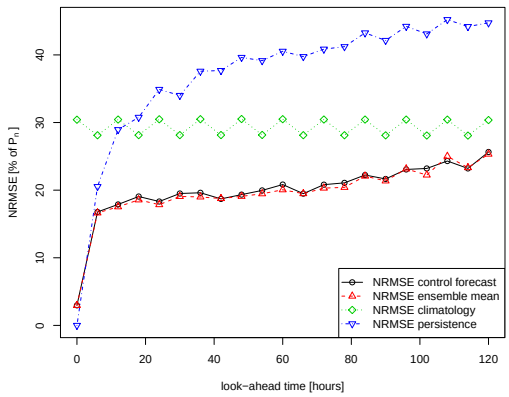
Appendix A

Evaluation of the forecasting models

A.1 NRMSE for Mardyck and Oupia using the RF model



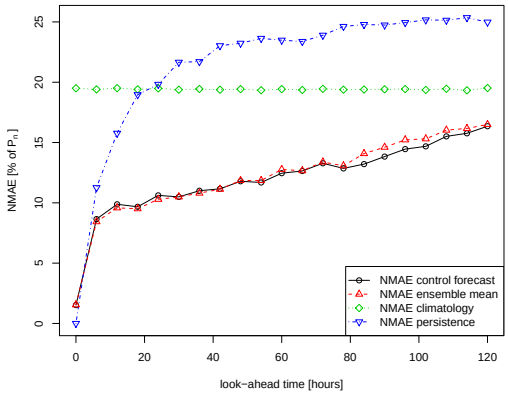
(a) Mardyck



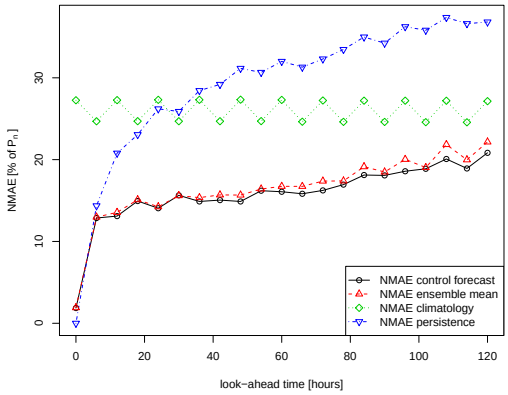
(b) Oupia

Figure A.1: NRMSE for Mardyck and Oupia for look-ahead times from 0 to 120 hours when the predictions are made with the RF model.

A.2 NMAE for Mardyck and Oupia using the RF model



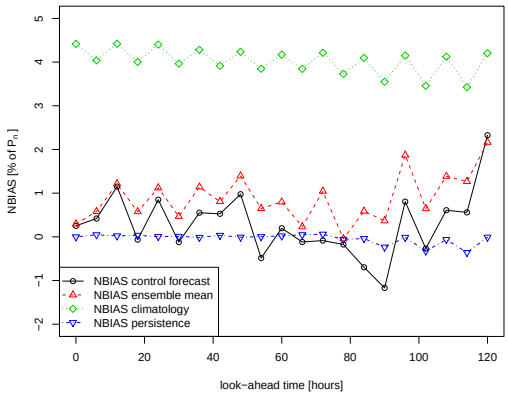
(a) Mardyck



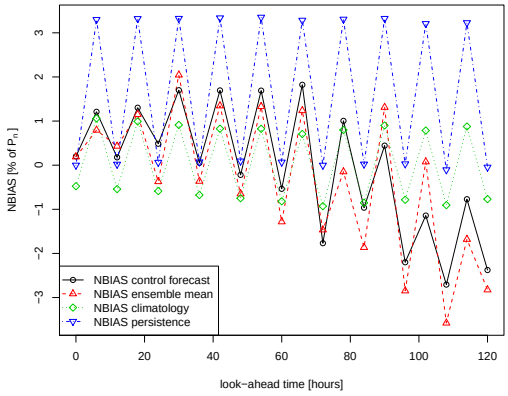
(b) Oupia

Figure A.2: NMAE for Mardyck and Oupia for look-ahead times from 0 to 120 hours when the predictions are made with the RF model.

A.3 NBIAS for Mardyck and Oupia using the RF model



(a) Mardyck

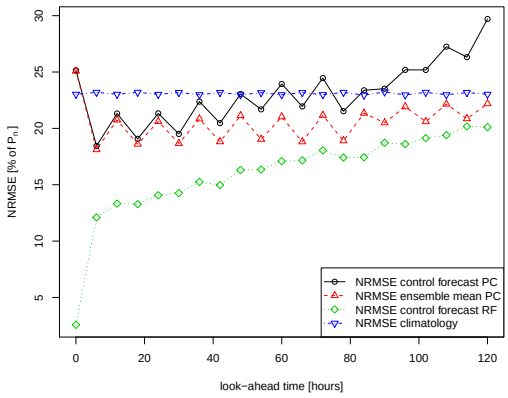


(b) Oupia

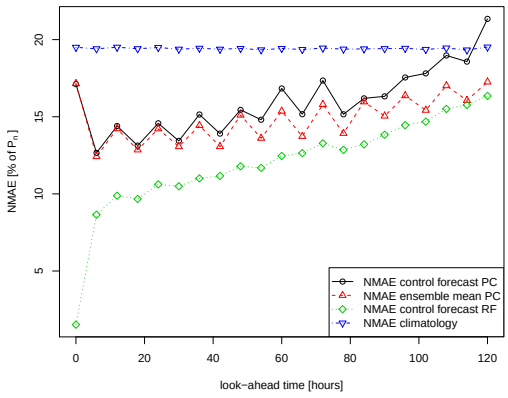
Figure A.3: NBIAS for Mardyck and Oupia for look-ahead times from 0 to 120 hours when the predictions are made with the RF model.

A.4 Comparison of NRMSE and NMAE for Mardyck using the PC model, the RF model and climatology

A.5 NRMSE and NMAE for the aggregate of the three wind farms

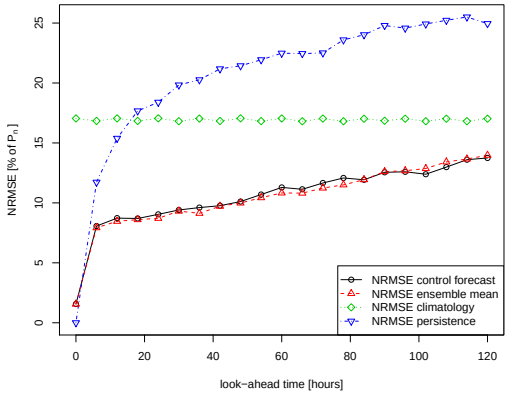


(a) NRMSE

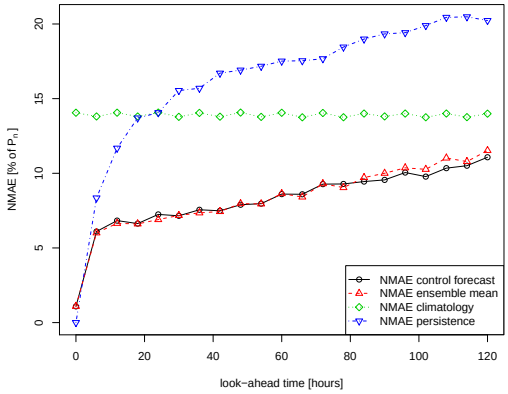


(b) NMAE

Figure A.4: NRMSE and NMAE for Mardyck using the Power Curve (PC) model, the Random Forest (RF) model and climatology.



(a) NRMSE



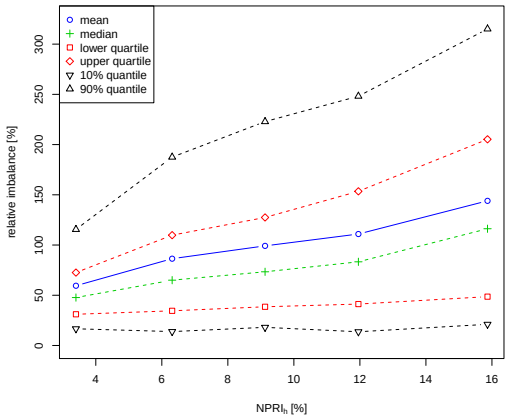
(b) NMAE

Figure A.5: NRMSE and NMAE for the aggregate of the three wind farms.

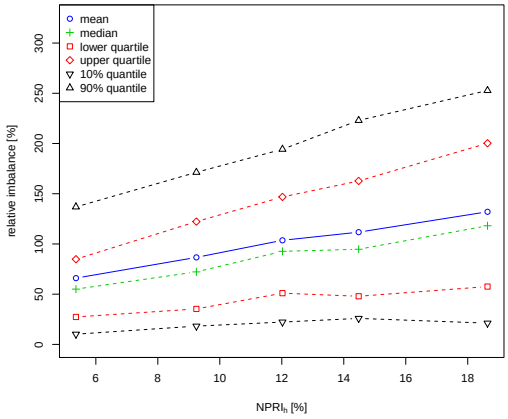
Appendix B

Analysis of the NPRI including extension to other temporal and spatial scales

B.1 $NPRI_h$ for Mardyck and Oupia using the RF model



(a) Mardyck

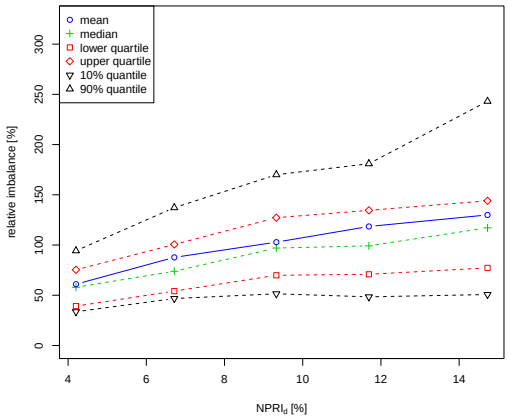


(b) Oupia

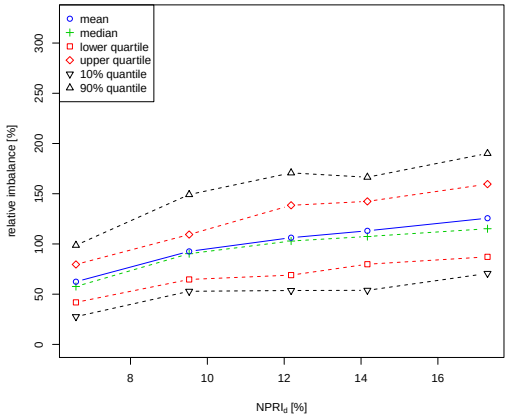
Figure B.1: $NPRI_h$ for day 3 for Mardyck and Oupia using ensemble predictions made with the RF model.

B.2 $NPRI_d$ for Mardyck and Oupia using the RF model

B.3 $NPRI_h$ of the power ensembles per look-ahead time for Mardyck and Oupia



(a) Mardyck

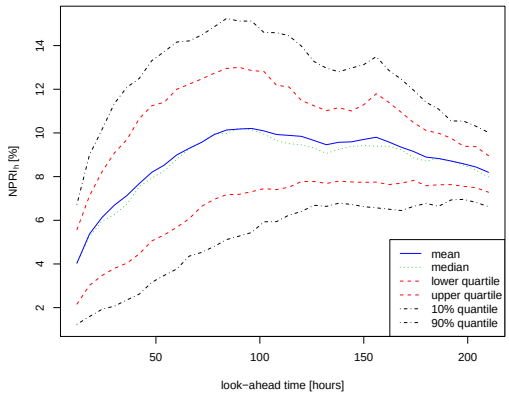


(b) Oupia

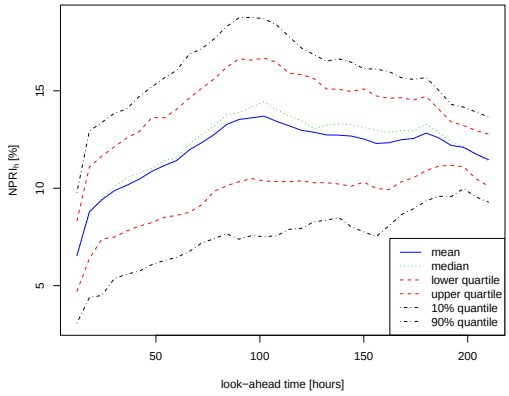
Figure B.2: $NPRI_d$ for day 3 for Mardyck and Oupia using ensemble predictions made with the RF model.

B.4 $NPRI_h$ of the wind speed ensembles per look-ahead time for Mardyck and Oupia

B.5 Statistics of the NPRI results using shorter and longer look-ahead time windows



(a) Mardyck



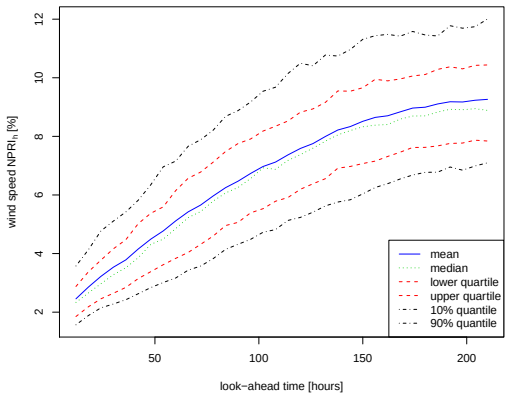
(b) Oupia

Figure B.3: $NPRI_h$ per look-ahead time for Mardyck and Oupia for ensemble predictions made with the RF model.

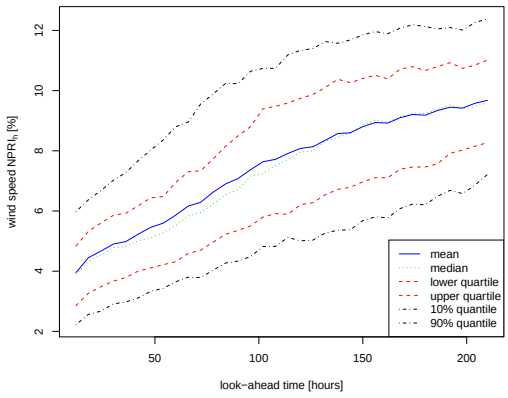
Table B.1: NPRI results using different lenghts of the look-ahead time window.

	Wind farm	$NPRI_d$ **24 - 72 h**	$NPRI_d$ **48 - 72 h**	$NPRI_d$ **48 - 60 h**
mean NPRI class 1-5	Mardyck	3.6 - 12.9	3.9 - 14.5	3.3 - 15.0
	Oupia	6.4 - 14.7	6.2 - 16.8	5.5 - 16.8
	Saint Simon	3.0 - 10.4	3.2 - 11.8	2.7 - 12.0
quote of mean imbalance level	Mardyck	2.12	2.22	2.23
	Oupia	1.65	1.77	2.01
	Saint Simon	2.17	2.30	2.22
minimum and maximum IQR	Mardyck	27 - 63	30 - 69	29 - 80
	Oupia	30 - 42	30 - 57	36 - 73
	Saint Simon	21 - 42	25 - 76	35 - 82

B.6 Results of $NPRI_h$ and $NPRI_d$ for day 1 for Saint Simon

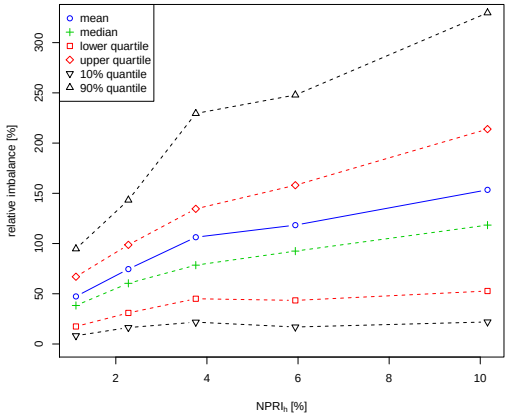


(a) Mardyck

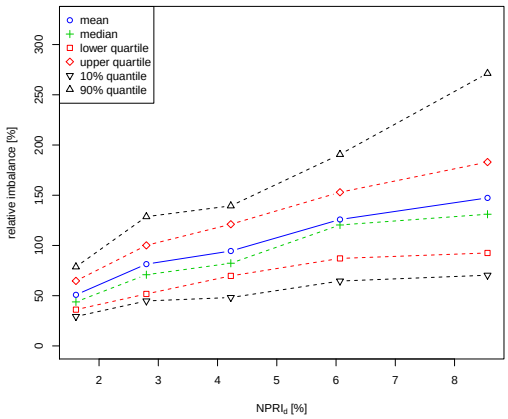


(b) Oupia

Figure B.4: Standard deviation of the predicted wind speed ensembles per look-ahead time for Mardyck and Oupia.



(a) $NPRI_h$



(b) $NPRI_d$

Figure B.5: $NPRI_h$ and $NPRI_d$ for day 1 for Saint Simon using ensemble predictions made with the RF model.

Appendix C

Alternative risk indices

C.1 Correlations between the energy imbalances obtained from the control forecast of the RF model predictions and the $NPRI_d$, the Max-min and the Max-min-max indices calculated on ensemble predictions generated by the PC model

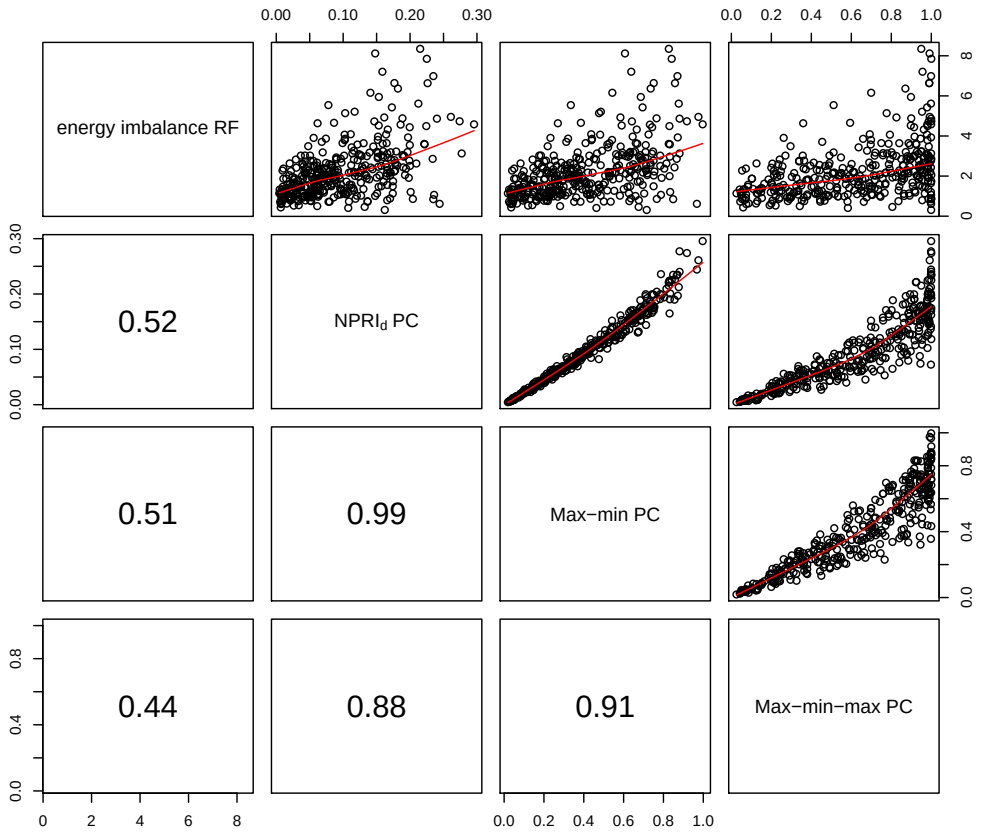


Figure C.1: Correlations between the NPRI and the two max-min indices for day 2 for Saint Simon when the indices are calculated on the predictions made by the Power Curve model and the energy imbalances are calculated on the predictions made by Random Forest.

C.2 Correlations between the energy imbalance calculated from the RF control forecast and the $NPRI_d$, the Willmott for horizons index, the Willmott for periods index and the Pearson Index of Agreement calculated from the ensemble predictions made with the PC model

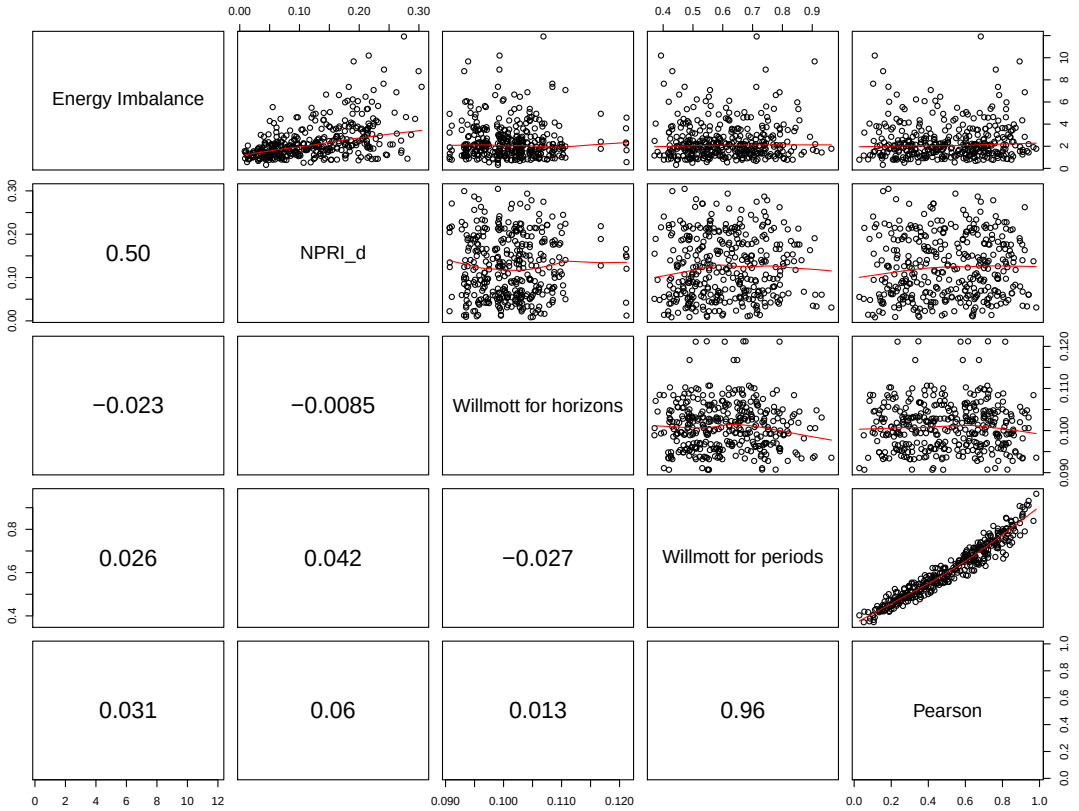


Figure C.2: Correlations between energy imbalance, $NPRI_d$, Willmott for horizons index, Willmott for periods index and Pearson index for day 3 for Saint Simon when the indices are calculated from the predictions made with the PC model and the energy imbalances are calculated from the RF control forecast.

Appendix D

Supplementary results

This chapter contains supplementary results from the work that were not relevant to include in the report. First of all, the observed deviations from a single power curve, when taking into account the number of active turbines, are presented. Furthermore, the choice of using all available data to training and testing is justified by comparing with the results of using date from the same period of the year. The decision to use the closest grid point to retrieve the NWP data, and not a combination of weighted mean of a number of points around the farm, is justified. Finally, the results using the skewness and kurtosis of the ensemble members as alternative risk indices are presented.

D.1 Deviations from a single power curve

In order to further investigate the properties of the measurements of wind speed and power, the measured power for a wind farm is plotted against measured wind speed. The plot for Mardyck, for which the wind speed measurements are obtained at hub height, is shown in Figure D.1. The points are drawn with different colours depending on the number of turbines that were active at the moment of the measurement.

A division of the power curve into several "branches" is seen for high wind speed. These branches correspond to different power curves originating from cases where one or several of the turbines are inactive. In this case, four different power curves are identified when 5, 4, 3 or 2 out of the 5 turbines are active. No cases where found where only one turbine was active. The differences between the different power curves become significant for larger wind speeds. There are however subdivisions into two or several power curves for all four branches. The reason for this is that among the five turbines, there are three different nominal powers. Combinations of three or four active turbines therefore give different total nominal power.

Another point of interest is that the power curve is not very sharp. This is probably due to the fact that average wind speeds are used, were the wind speeds are calculated as the mean of the measurements from the five turbines, together with the fact that no consideration is taken on the wind direction. Another remark is that there are power measurements that are larger than the nominal power. This is however not surprising since the nominal power is not necessarily equal to the maximum power.

The division into different number of power curves depending on the number of active turbines can be compensated for by normalizing the power output by the sum of nominal power for the active turbines. For simplicity, the nominal power is here considered to be the same for all turbines, equal to one fifth of the total nominal power of the farm. The instantaneous nominal power is therefore calculated as the nominal power times the ratio of the number of active turbines. The result is shown in Figure D.2.

The accordance with one power curve is now much better. There are however some large deviations, for example a couple of observations that are far higher than the nominal power. This indicates that the use of an actual nominal power explains part of the branching but not

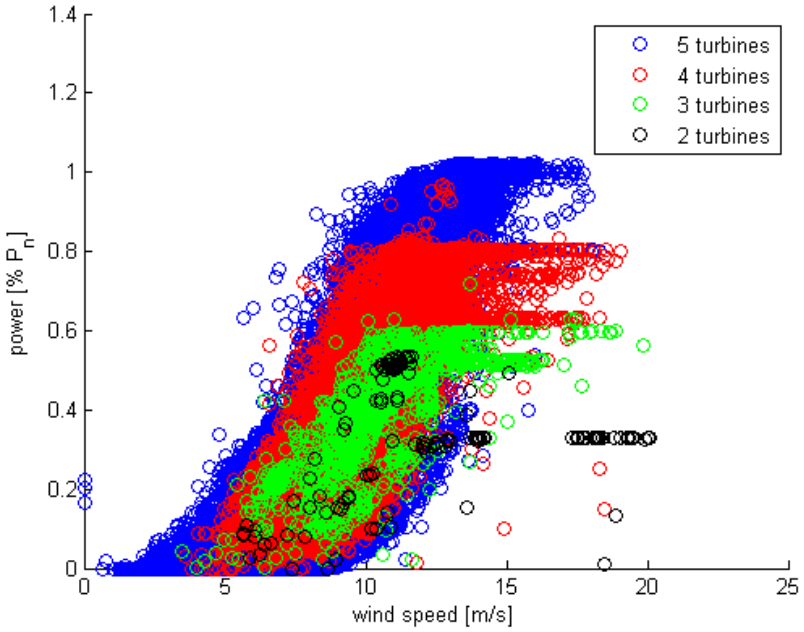


Figure D.1: Measured power as a function of measured wind speed for Mardyck. The points are separated depending on the number of active turbines.

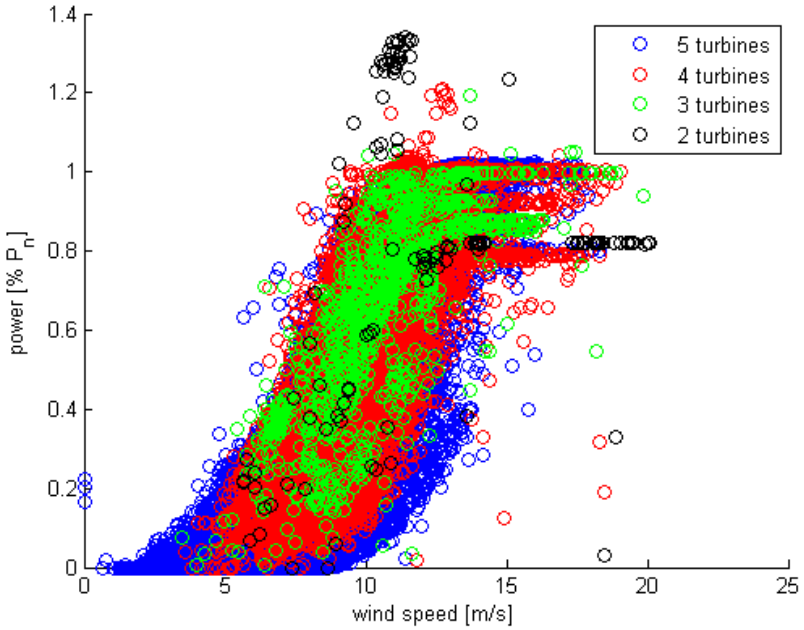


Figure D.2: Measured power as a function of measured wind speed for Mardyck. The power measurements are normalized by the nominal power times the fraction of number of active turbines.

everything. The remaining errors might occur from errors in the data, for example if a turbine is registered as active while it is not. It must also be remembered that the results would

have been better if the deviations in nominal power between the turbines had been taken into account.

These types of branches are less clearly seen for the other two farms which both have identical nominal power for all turbines. In the case of Saint Simon, the branching is very tenuous, indicating that all turbines have been active almost all the time. In the case of Oupia, there is a large spread among the power values for large wind speeds. The number of turbines is however larger, equal to 9, and this has for effect that the distinction between different power curves is smaller. The large spread can also be partly due to differences in power output depending on wind direction. This is more apparent for wind farms with more turbines due to shadowing effects as well as for wind farms located in complex terrain where the local wind conditions are more sensitive to the direction of the wind. Oupia corresponds to both these descriptions.

It could be interesting to use the number of active turbines as an explanatory variable in the predictions. The problem is then that predictions of the number of active turbines must be included. This could however be obtained in an operational context if the wind farm operator possesses information about the number of active turbines and scheduled maintenance. This option has however not been investigated in the frame of this work.

D.2 Predictions using the same periods for learning and testing

All of the results in the main report have been obtained using all the data from the 1.5 year with a learning set of the first 2/3 of the data. For comparison purposes, training on the 5.5 months period from first of June 2004 to 15th of December 2004 followed by a testing on the same 5.5 months on the year 2005 was carried out. Figure D.3 shows NRMSE for the control forecast for Saint Simon.

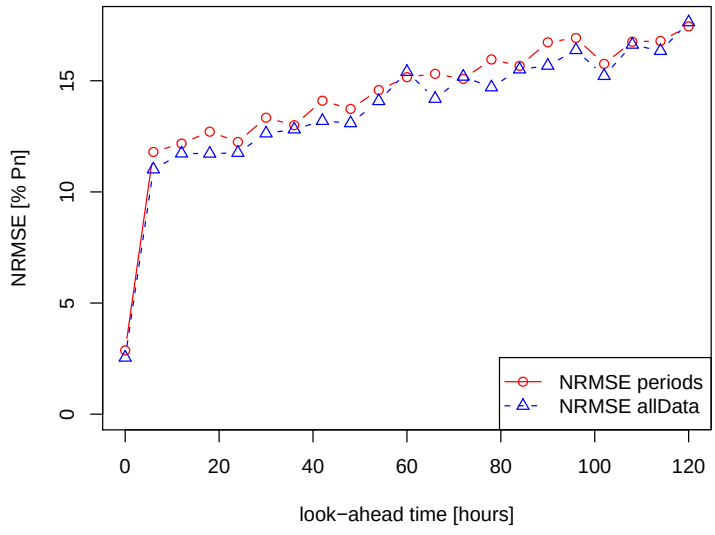


Figure D.3: NRMSE when using all data or shorter time periods for Saint Simon.

It is clearly seen that training and testing on the shorter time period increases the forecast error even though the training and testing take place on the same period of the year. This is probably due to the fact that more data is used for the model training when all data is used and that there is continuity between the learning and testing sets. A constant higher NRMSE

when using a shorter period of data is also obtained for the other two farms. The results are similar when considering NMAE. The conclusion of this is that all available data should be used to make as correct predictions as possible. This was also the decision taken in the main report.

D.3 Improvement using a weighted mean of surrounding points

In the main report, it was decided to extract the NWP data from the point in the grid that was situated closest to the wind farm. However, since the spatial resolution of the NWP data is rather coarse, 1^o in both longitude and latitude, the distance to the closest points for the three farms is 18.13 km for Mardyck, 37.46 km for Oupia and 30.99 km for Saint Simon. These distances are rather large and local effects can have large impacts and result in significantly different wind conditions between the point of the NWP grid and the farm. For Oupia for example, which is situated in complex terrain, this low resolution has for effect that the closest point is on flat terrain close to the Mediterranean Ocean. It is therefore motivated to consider alternative options.

D.3.1 Description of the alternative options

Instead of using the closest grid point, some kind of combination of the NWP values from the grid points around the wind farm can be used. The ideal would be to use a complete interpolation of the meteorological variables to a finer grid resolution. Due to the large amount of data with several dimensions; ensembles, look-ahead times and run times, this would however be a computationally costly option. Two easier alternatives are therefore considered; using a weighted mean of the 4 surrounding points and using a weighted mean of the n closest lying points. In both cases, the contributions of the points to the total have to be determined. An easy and reasonable way is to weight the points with the inverse of the distance d between the farm and the grid point, normalized by the sum of the inverse distances between the farm and the grid points:

$$w_i = \frac{1/d_i}{\sum_{i=1}^{n} 1/d_i} \tag{D.1}$$

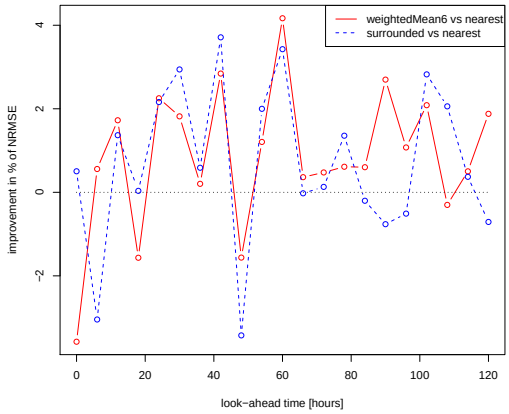
D.3.2 Impact on the forecast error

The reduction in forecast error by using some of these two more advanced and more computationally demanding methods is shown in Figure D.4 with a weighted mean of the 6 closest points. The results are shown as the improvement in % of NRMSE and NMAE for the Saint Simon wind farm.

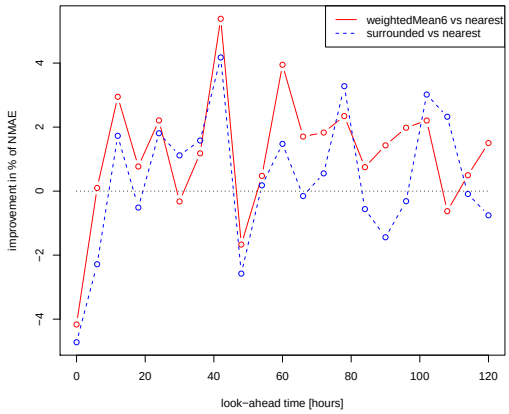
Considering NRMSE, it is observed that it is an advantage to use a weighted mean of the 6 closest points or the 4 surrounding points. The performance is however not positive for all look-ahead times and the average reduction in forecast error is small, 0.86 % and 0.70 % of NRMSE for the two options. For NMAE, the results are better for a weighted mean of 6 points and worse, but still in average positive, using the 4 surrounding points.

Small improvements using any of the two more advanced methods are also found for Oupia when considering NMAE. For NRMSE, there is an average reduction, caused by a large deterioration for look-ahead time 0 hours. For Mardyck, there are varying results but in total an improvement using the 6 closest points but deterioration using the 4 surrounding points. This can probably be explained by the fact that the latitude of the wind farm deviates very little from the nearest latitude in the grid, only by 0.008 degrees. The wind farm is therefore situated almost on the line between 2 grid points and using the 4 surrounding points might not be representable.

As another comparison, a weighted mean where the weights determined in other ways were studied for Saint Simon. It was shown that neither using the inverse of the square of the



(a) NRMSE



(b) NMAE

Figure D.4: Improvement in % of NRMSE and NMAE when using a weighted mean of the 6 closest points and a weighted mean of the 4 surrounding points compared with using the nearest lying point. The results are for the Saint Simon wind farm.

distances, $1/{d_i}^2$, nor a negative exponential of the distance, e^{-d_i}, gave better results than the inverse distances to the points. Since the inverse distances to the point is also easier and less costly to calculate, this option is to prefer.

To summarize it can be concluded that the results vary and the improvements found are minor, a reduction of 1 % of NMAE for Saint Simon only correspond to around 0.1 % of P_n. The average improvement also depends on how many look-ahead times that are taken into account. Making calculations based on the closest point is therefore considered to constitute a reasonable choice. Using the closest point is also the easiest and less computationally demanding option. In an operational context, it can also be easier and cheaper to retrieve NWP data from only one point then from a number of points. This justifies the choice of using the closest point in the work. It can however be interesting to compare with the weighted mean, for example when studying the NPRI.

D.3.3 NPRI results

Statistics for the NPRI when it is calculated for the closest point or a weighted mean of the 6 closest points for the three farms are shown in Table D.1. Results are for day 3.

Table D.1: NPRI results when using the closest grid point compared with a weighted mean of the 6 closest points.

	Farm name	$NPRI_h$ **closest**	$NPRI_d$ [%] **closest**	$NPRI_h$ [%] **weighted**	$NPRI_d$ [%] **weighted**
mean NPRI class 1-5	Mardyck	3.4 - 15.8	4.2 - 14.7	3.4 - 16.1	4.1 - 14.9
	Oupia	5.4 - 18.6	6.6 - 17.2	5.4 - 18.3	6.5 - 16.9
	Saint Simon	2.7 - 13.6	3.4 - 12.3	2.4 - 13.5	3.1 - 12.1
ratio of mean imbalance level	Mardyck	2.45	2.08	2.36	2.32
	Oupia	1.92	2.01	2.12	2.02
	Saint Simon	2.67	2.50	3.00	3.01
minimum and maximum IQR	Mardyck	43 - 160	33 - 69	47 - 151	32 - 67
	Oupia	58 - 136	35 - 71	47 - 131	34 - 61
	Saint Simon	41 - 167	23 - 97	40 - 162	23 - 106

A first remark is that the NPRI ranges are slightly larger for $NPRI_h$ than for $NPRI_d$, which is explained by the averaging of NPRI-values over the look-ahead time window for $NPRI_d$. These ranges do not change significantly when the weighted mean of the 6 closest points is used to compute the NPRI data.

Furthermore, it is found that using a weighted mean of the 6 closest points increase the ratios of mean imbalance level for Saint Simon from 2.67 to 3.00 and 2.50 to 3.01 for $NPRI_h$ and $NPRI_d$ respectively. These larger values mean that the usefulness of the index increases. Small changes are present for the other two farms, with an improvement in all cases except for $NPIR_h$ for Mardyck. This shows that even though the use of a weighted mean led to an average increase in forecast error over the first 5 days, which was the case for Mardyck, the NPRI can still distinguish better between low and high energy imbalances.

The IQRs are smaller for $NPRI_d$ than for $NPRI_h$ which is due to the averaging of NPRI values. Using a weighted mean gives generally a smaller IQR, and thus a better resolution and higher sharpness of the index. The improvements are however small and not systematic. There is also an increase in the IQR for the Saint Simon $NPRI_d$.

The conclusion of these varying results is that it can be advantageous to optimize the selection of NWP data points, using the closest point or a weighted mean of n closest points, in order to maximize the information that can be extracted from the NPRI. It is however important to determine on which criterion, lowest average NRMSE, highest NPRI resolution or other, the selection of NWP data should be made. Furthermore, the improvements must be viewed against the effort and computational cost of retrieving and assembling data from more grid points. In order to be consistent, it is therefore motivated to use the closest point.

D.4 Using the ensemble mean instead of the control member to calculate energy imbalances

In the report, the energy imbalances used for the risk indices have been based on the control member. An alternative to this is to calculate the imbalances on the ensemble mean. Even though the NRMSE and NMAE are similar for the control member and the ensemble mean, from observations of spaghetti plots it is clear that there can be large variations in the energy imbalances based on the two options. It is therefore of interest to investigate whether the NPRI better explains the energy imbalances from the ensemble average. Results for day 3 are shown for the three farms in Table D.2. The NPRI ranges are not shown since these are identical due to the fact that the NPRI is calculated in the same way in both cases.

Table D.2: NPRI results using energy imbalances from the ensemble mean or the control forecast.

	Farm name	$NPRI_h$ [%] **control**	$NPRI_d$ [%] **control**	$NPRI_h$ [%] **mean**	$NPRI_d$ [%] **mean**
ratio of mean imbalance level	Mardyck	2.63	2.67	2.51	2.55
	Oupia	2.20	2.04	2.14	1.95
	Saint Simon	2.68	2.48	2.57	2.34
minimum and maximum IQR	Mardyck	40 - 138	22 - 89	31 - 126	18 - 116
	Oupia	47 - 133	26 - 67	44 - 111	39 - 50
	Saint Simon	41 - 155	26 - 95	35 - 135	21 - 80

Using the energy imbalances calculated on the ensemble mean is found to give slightly lower ratios of mean imbalance level compared to when calculating energy imbalances on the control member. It could be partly explained by the fact that the ensemble mean was found to give a slightly higher NMAE than the control member in most cases. The IQRs are reduced for both the first and last class for all cases except class 5 for $NPRI_d$ for Mardyck. This higher sharpness means that uncertainties about the levels of energy imbalance in each class are smaller. This is an advantage and means that this manner of displaying the NPRI can be useful.

D.5 Skewness and kurtosis

Besides using the first two moments of the distribution of ensemble members, the mean and the standard deviation, it is also of interest to investigate whether the third and fourth moments, skewness and kurtosis, can be used as alternative risk indices to give supplementary information about the level of forecast error.

D.5.1 Standard ensembles

If the ensemble members are skewed this would indicate that the centre of gravity of the ensemble member distribution is not located at the same position as the mean. A large skewness would imply that some members are largely deviating on some side of the distribution.

A distribution with a zero kurtosis has the shape of a normal distribution. A positive kurtosis implies a sharper peak and lighter tails while a negative kurtosis implies a lower peak and fatter tails. In the context of ensembles, a large positive kurtosis would imply that a larger part of the ensemble members are concentrated around the mean value, indicating one main trajectory. A negative kurtosis on the other hand would mean that the ensemble members are more spread out, indicating several trajectories.

The study is carried out by calculating standard deviation, skewness and kurtosis on the ensemble member predictions for each look-ahead time and each run time separately. These are also compared with the energy imbalance calculated from the ensemble mean. This is made for each look-ahead time and each prediction time. Figure D.5 shows correlations between energy imbalance, standard deviation, skewness and kurtosis of the ensemble members. Points are shown for look-ahead times corresponding to day 2 and 3 for the Saint Simon wind farm.

There is a rather weak relation between the standard deviation of the ensemble members and the energy imbalance. The largest imbalances are found when the skewness of the ensemble members is small, corresponding to a well centred distribution. When regarding the relationship between kurtosis and energy imbalance it is found that the largest energy imbalances are present for negative or small kurtosis, which is situations where the ensemble members are spread out. There is however also many cases for which the energy imbalance is low even with a low skewness or low kurtosis. Furthermore, a non-linear relation between skewness and kurtosis is apparent. The relation shows that when there is a large skewness, there is also a large

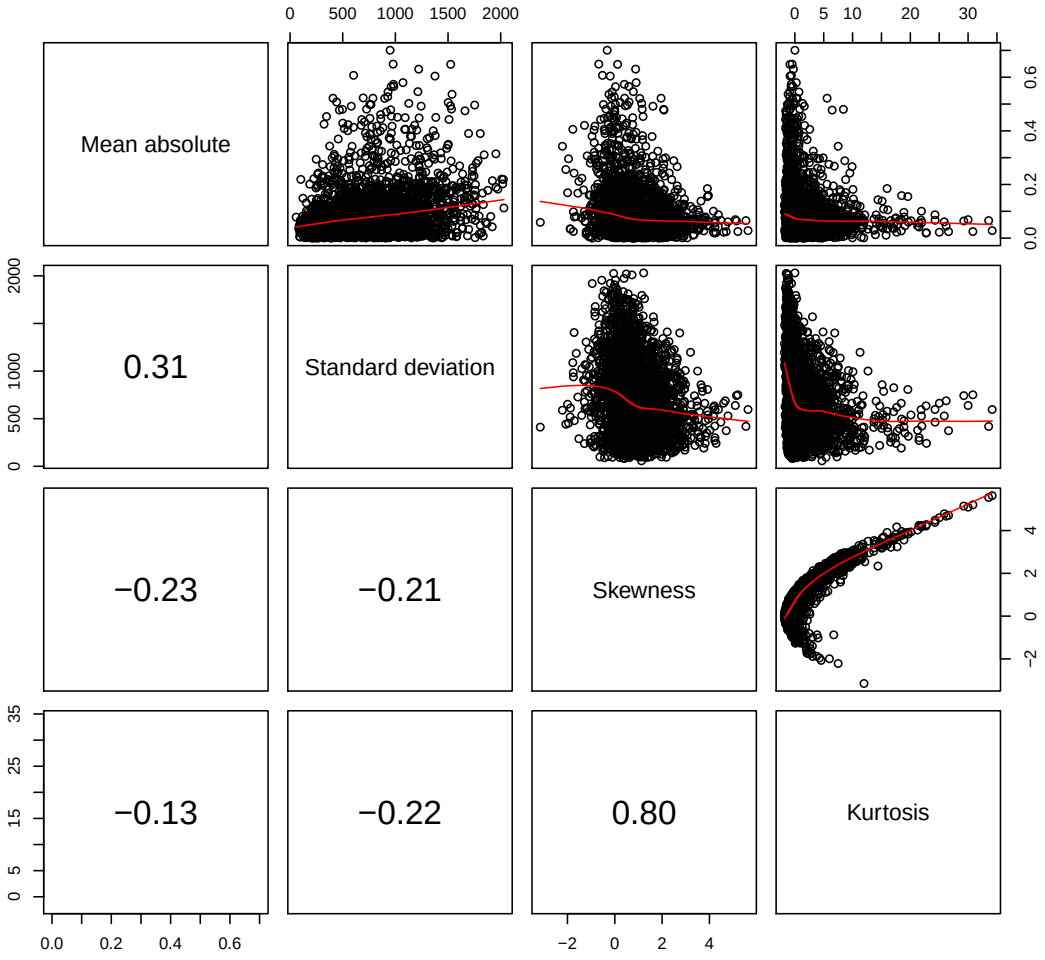


Figure D.5: Correlations between energy imbalance, standard deviation, skewness and kurtosis of the ensemble members. The results are shown for look-ahead times in day 2 and 3 for the Saint Simon wind farm.

kurtosis. In the same way, negative kurtosis is associated to low skewness. This is however not surprising since skewness and kurtosis both are based on the moments of a distribution and a relation between them can be explained mathematically.

D.5.2 Multi ensembles

One reason for the low correlations could be that there are only 51 ensemble members and that the ensemble members are not reliable in a probabilistic sense. In order to get more ensemble members, the same plot is drawn for the multi ensembles, shown in Figure D.6.

The correlations are more or less the same showing that the use of multi ensembles does not help to give more information on the energy imbalance. The correlation between skewness and kurtosis is even stronger, which is probably due to the fact that the statistics are calculated on a larger set of ensemble members.

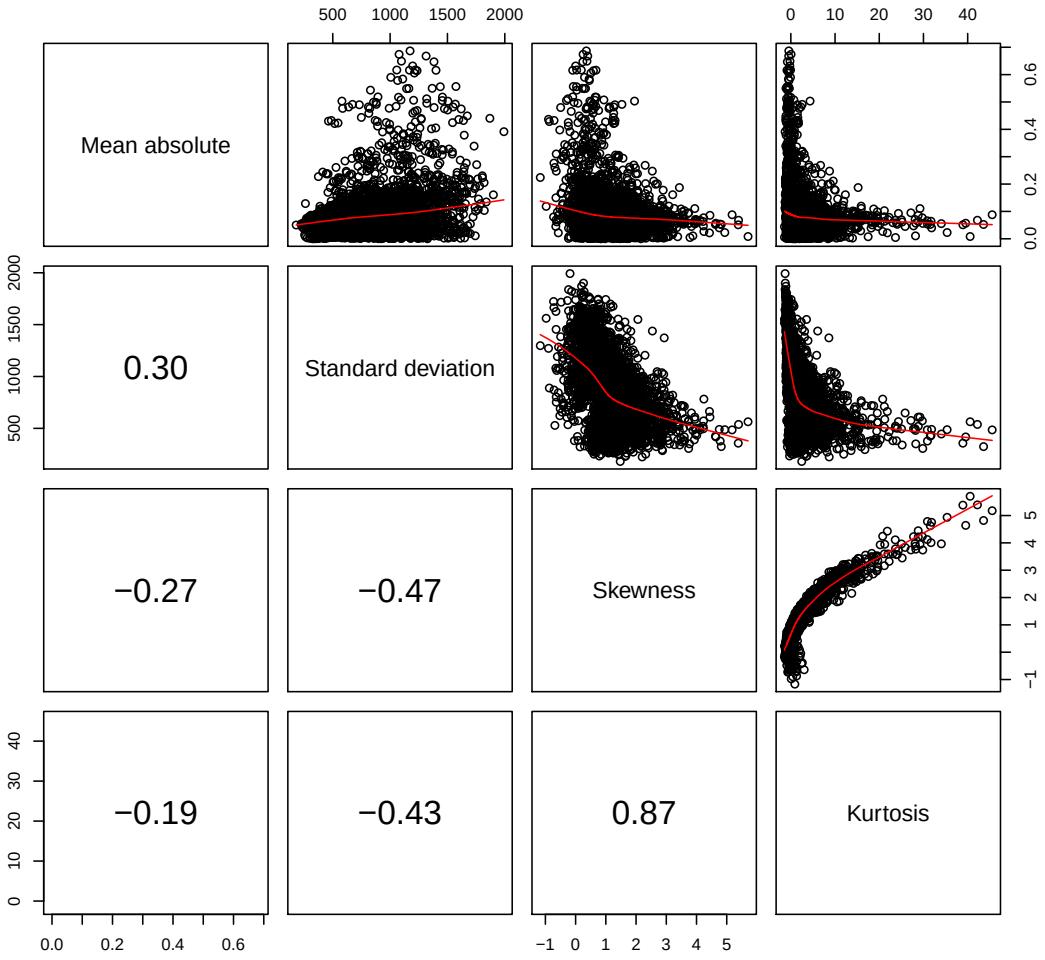


Figure D.6: Correlations between absolute error, standard deviation, skewness and kurtosis of the multi ensemble members. The results are shown for look-ahead times in day 2 and 3 for the Saint Simon wind farm.

Bibliography

[1] Calculate distance, bearing and more between two latitude/longitude points.

[2] Project anemos. Homepage.

[3] Project anemos.plus. Homepage.

[4] Project safewind. Homepage.

[5] *Uncertainty Estimation of Wind Power Forecasts*, 2008.

[6] Leo Breiman. Random forests. Technical report, Statistics Department, University of California Berkeley, 2001.

[7] Center for Energy and Processes, Ecole de Mines de Paris. *Advanced Strategies for Wind Power Trading in Short-Term Electricity Markets*, 2008.

[8] Martin Ehrendorfer. Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Zeitschrift*, N.F. 6:147–183, 1997.

[9] EWEA. *Wind Energy - The Facts*. EWEA, 2008.

[10] EWEA. Oceans of opportunity - harnessing Europe's largest domestic energy resource. Technical report, EWEA, 2009.

[11] Lionel Fugon. Prédiction de la production eolienne à partir de méthodes probabilistes. Technical report, MINES ParisTech, 2007.

[12] Gepsoft. Analyzing aps models statistically - confusion matrix, October 2009.

[13] Gregor Giebel. The state-of-the-art in short-term prediction of wind power. Technical report, Project ANEMOS, 2008. A Literature Overview.

[14] Gregor Giebel, Jake Badger, Lars Landberg, and Henrik Aalborg Nielsen. Wind power predictions using ensembles. Technical report, Riso National Labaratory, 2005.

[15] GWEC. Wind is a global power source. Homepage, July 2009.

[16] Thierry Jouhanique. Prédiction de la production Éolienne à partir de méthodes probabilistes. projet enseole. description des cas d'etude parcs Éoliens et données collectées. Technical report, EDF R&D - MFEE, 2007.

[17] Haleh Kootval. *Guidelines on Communicating Forecast Uncertainty*. World Meteorological Organization, 2008.

[18] H. Madsen, G. Kariniotakis, H.Aa. Nielsen, T.S. Nielsen, and P. Pinson. A protocol for standardizing the performance evaluation of short-term wind power prediction models. Technical report, Technical University of Denmark, Informatics and Mathematical Modelling. Ecole des Mines de Paris, Centre for Energy Studies, 2004.

[19] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[20] Henrik Aalborg Nielsen, Torben Skov Nielsen, Henrik Madsen, Gregor Giebel, Jake Badger, Lars Landberg, Kai Sattler, Lars Voulund, and John Tøfting. From wind ensembles to probabilistic information about future wind power production - results from an actual application. Technical report, Risø National Laboratory, Wind Energy Department Danish Meteorological Institution, Meteorological Research Division Energi E2 A/S Elsam Elkraft A/S, 2006.

[21] P. Pinson, H.Aa. Nielsen, H. Madsen, and G. Kariniotakis. Skill forecasting from ensemble predictions of wind power. *Applied Energy*, 86:1326–1334, 2008.

[22] Pierre Pinson. *Estimation of the Uncertainty in Wind Power Forecasting*. PhD thesis, Ecole des Mines de Paris, 2006.

[23] Pierre Pinson and Henrik Madsen. Ensemble-based probabilistic forecasting at Horns Rev. Technical report, DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark, 2008.

[24] Pierre Pinson and Henrik Madsen. Ensemble-based forecasting at Horns Rev: Ensembe conversion and kernel dressing. Technical report, DTU Informatics, Technical University of Denmark, 2009.

[25] Cort J. Willmott. Some comments on the evaluation of model performance. *American Meteorological Society*, 63:1309–1313, 1982.

[26] Cort J. Willmott, Steven G. Ackleson, Robert E. Davis, Johannes J. Feddema, Katherine M. Klinck, David R. Legates, James O'Donnell, and Clinton M. Rowe. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, 90:8995–9005, 1985.