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-5.8

"Skill Improvement of Extreme Event Forecast utilising a Calibrated Prediction System (CPS)"

DOCUMENT TYPE	Deliverable
DOCUMENT NAME:	swind.deliverable_Dp-5.8.doc
VERSION:	V1.0(*)
DATE:	2012.08.29
CLASSIFICATION:	R0: General Public
STATUS:	S0: approved

Abstract: The skill of ensemble forecasts as generated by the ECMWF integrated forecast system can be maximised by correcting for their lack of sufficient reliability. A bivariate calibration of these ensemble forecasts has been performed utilising adaptive and recursive estimation of the parameters of mean and variance models in a maximum likelihood framework. The originality of this methodology lies in the fact that calibrated ensembles still consist of a set of (space-time) trajectories, after translation and dilation. An adaptive calibration of ECMWF ensemble forecasts of (u, v)-wind at 10 metres above ground level was applied for Europe over a 3-year period between December 2006 and December 2009. Substantial improvements in (bivariate) reliability and in various deterministic/probabilistic scores were found concerning the full spectrum of wind events and wind power forecasts for Germany. In the short-range, the calibrated ensemble has a substantial positive impact to forecast extreme wind speeds and extreme wind power more precisely.

AUTHORS[1], REVIEWERS				
MAIN AUTHOR/EDITOR:	Lueder von Bremen			
AFFILIATION:	ForWind – Center for Wind Energy Research			
ADDRESS:	University of Oldenburg, Ammerländer Heerstrasse 136, D-26129 Oldenburg, Germany			
TEL.:	+49 441 798 5071			
EMAIL:	lueder.vonbremen@forwind.de			
FURTHER AUTHORS:	Thomas I. Petroliagis (ECMWF)			
PEER REVIEWERS:	Constantin Junk (ForWind, University of Oldenburg)			
REVIEW APPROVAL:	Approved by T. Petroliagis (WP-Leader)	29 August 2012	Rejected (improve as indicated below) :	
SUGGESTED IMPROVEMENTS:	*For a long list of remarks make reference to another document*			

VERSION HISTORY			
VERSION[2]:	DATE:	COMMENTS, CHANGES, STATUS:	PERSON(S):
v0.1	2011-08-01	First draft (joint work & results)	L.v. Bremen & T.I. Petroliagis
v0.2	2012-02-14	Second draft (additional tests & results)	T.I. Petroliagis
v0.3	2012-08-09	Third draft with CPS results for wind power forecasting (for review)	L. v. Bremen
v0.7	2012-08-20	Revised version after peer-review	L. v. Bremen
v1.0	2012-08-29	Approved by WP-Leader	T. Petroliagis

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 and the European Community is not responsible for any use that might be made of data appearing therein.**

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

Contents

Nomenclature

NOMENCLATURE

NWP:	Numerical Weather Prediction
ECMWF:	European Centre for Medium-Range Forecasts
IFS:	ECMWF Integrated Forecast System (Deterministic Platform)
EPS:	ECMWF Ensemble Prediction System (Probabilistic Platform)
VarEPS:	ECMWF Variable Resolution EPS (extension of EPS)
SV:	Singular Vector (method)
BV:	Bred Vector (method)
BMA:	Bayesian Model Averaging (calibration)
CRPSS :	Continuous Rank Probability Skill Score
BSS :	Brier Skill Score
NCEP:	National Centres for Environmental Prediction
NAEFS:	North American Ensemble Forecast System
WPF/P:	Wind Power Forecasting/Prediction
GEPS:	Global Ensemble Prediction Systems
LEPS / REPS:	Limited-Area / Regional Ensemble Prediction System
PDF:	Probability Density Function
ECFS:	ECMWF File Storage System
ERA-40:	ECMWF 40-year Re-analysis Project
ERA-Interim:	ECMWF Reanalysis Project for 1989 to 2013
DWD:	German National Meteorological Service
M-Range:	Forecast horizon of Medium-Range (120 to 168 hours) also denoted as *M-R*
Early M-Range:	Forecast horizon of early M-R (60 to 120 hours) also denoted as *early M-R*
Late M-Range:	Forecast horizon of late M-R (168 to 240 hours) also denoted as *late M-R*
S-Range:	Forecast horizon of Short-Range (12 to 60 hours) also denoted as *S-R*
Very S-Range:	Forecast horizon of very S-R (0 to 12 hours) also denoted as *very S-R*
Nowcasting:	Forecast horizon of 0 to 6 hours (being included in the very S-R)
TSOs:	Transmission System Operators
WMO:	World Meteorological Organisation
GEM:	Global Environmental Multi-scale (model) of CMC
OI:	Optimum Interpolation
UK Met Office:	Meteorological Service of U.K.
Meteo France:	Meteorological Service of France
AEMET:	Meteorological Agency of Spain
GLAMEPS:	Grand Limited Area Model Ensemble Prediction System
VSREF:	Very Short Range Ensemble Forecast System
UM:	Unified Model (UK Met Office)
SREPS:	Short Range Ensemble Prediction System

1. Introduction

Ensemble forecasts of (u, v)-wind are of crucial importance for a number of decision-making problems in operational meteorology and in particular for wind power forecasting [1]. To provide better predictions, some meteorological centres conduct statistical post processing to their EPS products for bias correction and calibration of the PDF (Probability Density Function). Bias correction is the most widely applied procedure, and an adaptive algorithm with Kalman Filter type weighting functions and the use of reforecast data make it more effective. At some weather centres, combination of an EPS with the corresponding high resolution deterministic prediction is operationally implemented, and statistical downscaling with high resolution analysis as the reference is used to provide forecast guidance at local scale [2]. The Bayesian Model Averaging (BMA) technique developed at the University of Washington is becoming quite widely used and is particularly well-suited to systems incorporating multiple models or parameterizations [3,4]. Some more sophisticated techniques for calibrating the first and second moments of the PDF are also under development.

Nevertheless, it is becoming clear that the skill of raw ensemble forecasts as generated by NWP-based models can be maximised by correcting for their lack of sufficient reliability. The original framework developed at ECMWF within SafeWind [1] allows for an adaptive bivariate calibration of these ensemble forecasts. The originality of this methodology lies in the fact that calibrated ensembles still consist of a set of (space-time) trajectories, after translation and dilation. In parallel, the parameters of the models employed for improving the stochastic properties of the generating processes involved are adaptively and recursively estimated to accommodate smooth changes in the process characteristics and to lower computational costs.

This approach is applied and evaluated based on the adaptive calibration of ECMWF ensemble forecasts of (u, v)-wind at 10 m above ground level over Europe over a 3-year period between December 2006 and December 2009. Substantial improvements in (bivariate) reliability and in various deterministic/probabilistic scores are observed over the full spectrum of wind events. Further details can be found in SafeWind's combined Deliverable 5.6/5.7 [5] that describes the developed Calibrated Prediction System (CPS) for ensembles.

The current work is focusing on extremes as the SafeWind project puts special emphasis on such events to develop together with grid operators the best solutions to integrate wind power into the grid. It is important to recall that one of the most important tasks of National Meteorological Services is to help forewarn society about severe or high-impact events that can result in considerable damage and large losses [6]. Much of the benefit for society through improved weather forecasts will come from advances in our capability to forecast such events so that mitigating actions can be taken.

Severe events are usually considered to be rare events, hence, the use of the term 'Rare Severe Event' (RSE) by Murphy [7]. Such events are also loosely referred to as 'Extreme Events' in atmospheric science [8]. Extreme events can come in many forms, such as intense multi-cell thunderstorms, tropical and extra-tropical cyclones, very intense wind events, heavy rain events, extreme heat and cold, floods and droughts. In our case, emphasis is given to windstorms, wind extremes and high wind power penetrations, since they are of great importance for save wind power integration into the power supply system.

In this Deliverable, the predictability of extreme events and how to deal with extremes are described in Section 2, while the concept of calibration in general and of CPS is presented in Section 3. Section 4 contains basic information about the technical implementation of CPS. The skill assessment of raw and calibrated ensembles is presented in Section 5 utilizing reliability diagrams. Talagrand diagrams of the raw and calibrated ensemble are shown in Section 6 for all events while in Section 7 the emphasis is on extreme wind events. Section 8 demonstrates the improvement of wind power forecasts for Germany by calibrated ensembles. Results are discussed in Section 9, while an Executive Summary is given in Section 10.

2. Forecasting extreme events

Extreme events are of great interest to the SafeWind project, since SafeWind focuses on the early detection and forecasting of such extreme events. Furthermore, it is obvious that extremes pose a special problem because they are *infrequent, poorly documented by observations, and at the limit of predictability.* Quantitative verifications of extremes are therefore more difficult and their statistical significance is mostly poor. Furthermore, besides a tolerance on space and time, a tolerance on the value of weather-related parameters should often be accepted in the case of extreme values. At the same time, it is recognized that even a poor numerical forecast in absolute terms can be of great value if it is well interpreted by an experienced forecaster.

The issue of extremes is made more complex by the scale difference between model and observations. In many cases one should not expect that current models reproduce the maximum values of weather parameters observed in extreme events because of representation errors due to relatively low resolution. We should however design methods to diagnose severe weather based on the existing models, and thoroughly verify the validity of these diagnostics [9].

2.1 Predictability of extreme events

In operational forecasting, a “gap” seems to exist between the events for which forecasters need to issue warnings and alerts, and what the numerical model guidance can provide to the forecasters. Some types of weather responsible for damage (lightning, wind gusts at different heights, fog) may not be explicitly simulated (predicted) by the model, and must therefore be diagnosed from other variables. Even if they can be explicitly predicted (e.g., heavy rain), the model resolution may not capture the intensity of the local weather and the processes associated with the variable are often at sub-grid scale. Some mesoscale models are being run experimentally at resolutions of 1-2 km, but most operational mesoscale models have grid scales of 5-15 km, while global models are still coarser.

Nowadays, warnings and alerts are often issued as areas with medium or high risk (probabilities exceeding certain predefined critical values) of experiencing a particular type of high impact weather. In the short range (0.5 to 2.5 days) and medium range (3 to 7 days) NWP ensembles are used to generate probability forecasts of the occurrence of extreme temperature, heavy rain, strong winds, and other high impact events. Verification of probability forecasts requires many matched forecasts and observations. This may be difficult to achieve for high impact weather which is often rare by definition. Verification using only a small dataset leads to results with large statistical uncertainties. Nevertheless, ensemble verification requires the same amount of data as any deterministic forecast.

Studying extreme wind events (such as windstorms) of the past, it is interesting to note that only a small proportion of ensemble members (or of deterministic forecasts from different weather centers) succeeded in predicting severe storms, even ~24 h in advance. It is important to keep in mind that in a synoptic situation when severe weather could occur, most ensemble members are likely to be drawn towards the model’s climatology once a forecast moves into the chaotic non-linear regime [10]. Consequently, it occurs that the control forecast and some perturbed members predicts severe weather and most perturbed members lead to less severe conditions.

Based on this, the forecast PDF is always likely to be skewed away from severe weather, i.e. the PDF is not centered on the severe event. Thus, although the ensemble can be expected to include members with severe events, it would be unusual for the forecast system to predict high probabilities of severe weather. Since the above analysis applies equally well to the real atmosphere as to a model, it can be argued that the occurrence of severe weather is fundamentally a low probability event in the atmosphere (in the medium-range) and that on most occasions it should be appropriate to issue early warnings at low probabilities.

2.2 Dealing with extremes

Common sense seems to dictate the use of very high-resolution models in order to cope with extreme weather events. It is true that the higher the resolution is, the wider the range of weather phenomena that can be described explicitly, including potentially damaging ones such as squall lines, tornadoes or tropical cyclones. It is now widely recognised, however, that not all phenomena that can be represented with realism in high-resolution models are predictable. This is because of the very demanding error constraints on the prescription of the initial state that cannot be achieved with the observing systems currently in operations [11]. On the synoptic scales, optimal control theory offers a convenient framework where the sensitivity of the forecast to small errors in the initial conditions can be demonstrated using adjoint models [12, 13]. Furthermore, the Fronts and Atlantic Storm-Track EXperiment (FASTEX) has been a large scale cooperative effort to gather for a two months period the supplementary data needed to improve the reliability of deterministic forecasts of rapidly developing perturbations over the Northern Atlantic [14].

The ability of current NWP platforms (such as the ECMWF IFS & EPS) to capture severe storms has improved in recent years, although the direct comparison of model wind speed over land with observations shows a fairly large negative bias. Among the reasons why this occurs is that modellers have tended to concentrate on the need to have a good momentum budget rather than on the objective of optimising on-site validation of local effects during the design of boundary-layer parameterisations. A step towards improved post-processing of maximum wind gust values based both on explicitly predicted model winds and on the sub grid scale representation of turbulent fluxes was, however, achieved in the year 2000 at ECMWF, resulting in a better correspondence between model predictions and observations. Furthermore, early warnings for extreme weather conditions such as wind gusts can be extracted from the EPS. The 51 EPS forecasts can be used to predict the probability that a particular weather event of interest (such as extreme wind speeds) might occur. The EPS provides an estimate of the likelihood of such an extreme event, given the inherent uncertainties mentioned above. As for the Extreme Forecast Index (EFI), the more the EPS distribution departs from the climatological curve the higher the EFI (in absolute value).

Furthermore, inference of skill for higher threshold forecasts *by issuing and assessing forecasts at lower thresholds* can provide forecasters with useful experience and feedback in forecasting and interpreting extreme events. For example, to have sufficient numbers of events for longer range forecasting systems such as current operational seasonal forecasting systems, forecasts are issued for (moderate) extreme events that are not very rare: tercile categories defined by the 0.15 and 0.85 empirical cumulative probabilities have been used at the Met Office and ECMWF to define such events as extremes in summer mean temperatures. It is often the case that deterministic forecasts are hedged to avoid missing warnings of severe events which leads to a large frequency bias (false alarms). For extreme events it is therefore desirable to issue probability forecasts that cannot be as easily hedged [7].

It should be noted that signal detection of severe weather events in the medium-range is likely to be difficult. It should be commonly agreed that *issuing and communicating warnings* at medium- and late medium-range based on predicted intense cyclones and/or synoptic weather types linked to destructive wind events should not be considered as an automated warning system. The impact of warnings with lead times of 5-10 days is rather a “warning light” that ensures that a potentially dangerous event is not unnoticed by the forecasters and TSOs.

Focusing on extremes, emphasis should be given on the predictability of severe weather in the short- and early medium-range considering *issuing and communicating early warnings*. There have been many cases in the past that could serve as bias estimators of severe weather forecast performance: in many cases it was possible to identify a signal that the forecaster should in retrospect had been aware of [15]. By contrast, conducting a verification study using forecaster’s expertise in real time or delayed mode is both costly in terms of human resources and biased in its own way by the forecaster’s perspective. Medium-range forecasts are hardly ever looked at in the context of severe weather.

3. The concept of calibration

The significant advances in the development of ensemble forecasting methodologies [16] & [17]and more generally probabilistic forecasts of meteorological variables have increasingly been considered as a crucial input to a number of socially-relevant decision-making problems. With increasing shares of wind power worldwide, probabilistic forecasts of near-surface and hub height winds are becoming increasingly popular. This is partly owing to the needs for accurate forecasting of wind power generation (from the short to medium range). A review of methods for ensemble-based forecasting of wind power is given in [18], [19] & [20]. It has been shown that the optimal management and trading of wind energy generation calls for probabilistic forecasts, see [21], [22] & [23] among others. From a more general point of view, probabilistic forecasts of near-surface winds and near hub height winds can be of great value for decision-making problems related to sailing, ship routing, air traffic control, etc. This statement is supported by theoretical results that for a large class of decision-making problems, optimal decisions directly relate to quantiles of conditional predictive densities [24].

As it is often the case for forecasts directly taken as output from physical models, ensemble forecasts of near-surface winds tend to be biased. For probabilistic forecasts this deficiency consists of their lack of sufficient (probabilistic) reliability: they are generally under-dispersive. With that in mind, various approaches to the bias-correction and calibration of ensemble forecasts of wind speed [4] & [25] and direction [2] have been described. The question of the bivariate view of wind speed and direction was in parallel touched [26] when discussing the skill evaluation of multivariate probabilistic forecasts. The prime aim here is to further develop on the calibration of ensemble forecasts of (u,v)-wind components in a multivariate framework. In contrast with the argument of Wilks [27], we are not looking at fitting probability distributions based on the ensembles.

The output of our calibration methodology consists of ensemble forecasts similar in nature to the uncalibrated ones, though with improved statistical properties. This approach is motivated by the fact that a number of decision-support systems using wind probabilistic forecasts as input need ensembles (in other words, trajectories) instead of predictive densities [28] & [29]. Indeed by fitting probability distributions for each point in space and in time, individually, the spatio-temporal structure of ensemble members gets lost.

The calibration methodology developed at ECMWF within SafeWind [1] & [5] is mainly inspired by the approach described in [19] for the adaptive kernel dressing of ensemble forecasts in a univariate framework, but also by the ideas described in [3] for probabilistic forecasting of (u, v)-wind using Bayesian Model Averaging (BMA). The present proposal is developed in a multivariate Gaussian framework, with the idea of correcting the first and second order moment properties of the ensemble forecasts. The main innovations brought in by this approach are that (i) u and v wind components are jointly considered instead of focusing on wind speed and potentially direction, individually, (ii) the output of the models are ensemble forecasts of the same nature than the input ones, not predictive densities, and (iii) the model parameters are considered as time-varying, while being adaptively and recursively estimated in a rigorous Maximum-Likelihood (ML) framework.

3.1 Calibrated Prediction System (CPS)

3.1.1. Rationale of CPS calibration

Proposals for the calibration of ensemble forecasts can already be found in the work in [27] & [30] among others, based on fitting probability distributions to the set of ensemble members. Since bivariate Gaussian variables are fully characterised by their mean vector and covariance matrix, fitting appropriate probability distributions would translate to the estimation of their mean and covariance for instance in a maximum likelihood framework. This would comprise a generalisation of the univariate case, as considered in [31] for the specific case of wind speed and in [32] a more general set-up.

Since aiming at conserving the original nature of the ensemble forecasts, a little twist has been introduced to these approaches by avoiding the direct fitting of distributions. The proposal is instead to concentrate on the underlying generating processes for the ensemble forecasts and for the error of the ensemble mean for every lead time, in order to introduce a two-dimensional translation and dilation of the sets of ensemble forecasts. This reduces to propose models for the mean and variance of the bivariate Gaussian densities, then yielding translation and dilation factors.

The translation corresponds to the (bivariate) bias-correction of the ensemble mean, while the dilation translates to the variance correction of the (unbiased) ensemble forecasts along the u and v dimension. The potential correction of the (u, v)- correlation is not considered since it would be difficult to apply it without having to resample from the generating processes, which is exactly what was aimed to be avoided. Similarly to [28] it has been considered that original data may be transformed before applying to place us within a bivariate Gaussian framework with linear models for the mean and variance. Using such a transformation was not deemed necessary in the present study based on ECMWF data. Studying the interest of considering more advanced models for the calibration of ensemble forecasts of (u, v)-wind may be the topic of further research.

Details and graphical illustration of a two-dimensional translation and dilation of the sets of ensemble forecasts can be found in [5]. An important observation from the empirical work and the maps of translation and dilation factors is that model parameters may certainly be represented by a spatial model, since exhibiting smooth variations in space. The simplification resulting from using a spatial model would also contribute to lowering computational costs.

3.1.2. Theoretical background of CPS

An exploratory analysis indicated that a set of linear models would be sufficient for the calibration of the ensemble forecasts of (u, v)-wind. Based on such parametric assumptions for the generating processes of forecasts and errors of the ensemble mean, it is proposed to estimate the parameters of the models through a Maximum Likelihood (ML) approach. It is thus aimed at maximising the likelihood of the observed wind vectors, given the calibrated probabilistic forecasts resulting from the model. More specifically, the method is a Recursive Maximum Likelihood (RML) approach, with exponential forgetting of past observations. An advantage of such a proposal is that only the last available set of forecasts and measurements (analysis) is employed at a given time t for updating the model parameters.

It hence allows for significant lowering of computational costs compared to the more traditional batch estimation methods, e.g. using a moving window of 3 months for estimating model coefficients. Another advantage brought in by the exponential forgetting is the ability for the model parameters to smoothly evolve, as a reaction to changes in the joint forecasts-observations process characteristics. These changes may originate from changes in the wind dynamics e.g. due to seasonalities, but also from changes in the forecasting system, like at the occasion of a change of model physics or of a change of horizontal/vertical resolution.

4. Implementation of the Calibrated Prediction System (CPS)

4.1 Raw probabilities by ECMWF EPS

The ECMWF deterministic IFS and probabilistic EPS platforms provide useful deterministic forecast guidance (IFS) and information on uncertainty (EPS) of the weather prediction for many different end-users and the general public. In this sense it is critical that ensemble forecasts represent a “reliable” distribution of probabilities that can develop from an initializing set of analysis fields. A probabilistic forecast is reliable if the frequency of occurrence of a certain event equals the forecasted probability that this event occur. It is important that these probabilities can provide early warnings in cases of extreme wind events, since the ultimate value of a weather forecast system lies in its ability to improve weather-related decision-making processes. The ECMWF IFS [33] has currently a horizontal resolution of ~16 km (T1279) with a 91 vertical level scheme, while the EPS has a horizontal resolution of ~32 km and a vertical resolution of 62 levels. The EPS was implemented operationally 20 years ago [34] & [35] and has undergone many changes since then [36]. In practical terms, the value of an ensemble prediction system is that it gives forecasters the means to access quantitatively their risk to weather sensitive events some days in advance. The current EPS comprises 50+1 members made with a T639L62 model to D+10 (max horizon of 240 hours), changing to a T319L62 formulation from D+10 to D+15 (maximum forecast horizon of 360 hours). The probability of a given event is determined from the fraction of ensemble members, which predict the event.

For the CPS implementation the ensemble forecasts of (u, v)-wind at 10 metres above ground level originate from the operational ensemble forecasting system at ECMWF. The forecast length considered is 5 days, corresponding to the lead times of interest for most of the decision-making problems involving wind forecasts. The domain chosen for this study is Europe (10°W- 23°E and 35°N-58°N). This domain covers a rectangular latitude-longitude grid with S = 80×57 = 4560 grid nodes. Future work may consider the possibility of evaluating the approach proposed in the present paper over the whole globe, in order to assess its interest under various climates. A smaller domain corresponding to the greater area of Germany (labelled: Germany) was also used for the evaluation of skill comprising 504 grid nodes.

Data including ensemble forecasts and the related model analysis from ECMWF has been collected over a period spanning December 2006 - December 2009. These ensemble forecasts are issued twice a day at 00 UTC and 12 UTC, with a horizontal resolution of about 50 km (corresponding to a spectral truncation at wave number 399). Note, that the available resolution is substantially coarser than during the time of writing (see above). The temporal resolution is 3 hours. But, since the model analysis, which is seen as a reference has a temporal resolution of 6 hours only, we consider this coarser resolution in the present study. The methodology employed for the generation of the ECMWF ensemble forecasts is well documented and a number of publications can be pointed at for its various components. For a general overview, see [36]. It is not the objective of this work to discuss and analyse competing methodologies for the generation of ensemble forecasts or more generally of probabilistic forecasts of meteorological variables. A comparison with other global ensemble prediction systems can be found in [37].

The ECMWF ensemble predictions aim at representing uncertainties in both the knowledge of the initial state of the atmosphere and in the physical parametrisation of the numerical model used for integrating these initial conditions. For the former uncertainties singular vectors are employed, the core methodology being extensively described by Leutbecher and Palmer [38]. A comparison of the different methodologies for the generation of initial perturbations can be found in [39]. In parallel for the latter type of uncertainties, stochastic physics is employed for sampling uncertainties in the parametrisation of the numerical model [40] & [41]. Note that the potential structural model uncertainty is therefore not accounted for.

4.2 Calibrated probabilities by CPS

From the available data, two periods are defined, the first one being used for identification (and initial training) of the statistical models, and the second one for evaluating the performance of these models under operational conditions. The first year of data is employed as the learning set, exactly covering the months from December 2006 to November 2007. The remainder of the dataset, covering a period from December 2007 to November 2009 is used for out-of-sample evaluation of the reliability and skill of the ensemble forecasts of (u, v)-wind, before and after calibration. We do not use the month of December 2009 for the out-of-sample forecast evaluation since focusing on complete (seasonal) quarters only.

Over the learning period, a part of the data is used for one-fold cross validation (the last 6 months), in order to select an optimal forgetting factor for the various models involved in the translation and dilation of the ensemble forecasts. Actually, instead of considering the forgetting factor itself, it is preferred to use the corresponding effective number of observations $n\lambda = 1/(1 - \lambda)$. It allows to better appraise the size of the equivalent 'sliding window' in the adaptive estimation of the dynamic model parameters, such as that considered in [12] and [42] for instance.

The selection of optimal values for the model structure and parameters $n\lambda$ is done in a trial-and-error manner, by evaluating the results obtained from a set of different setups. For more information on cross validation, it is referred to [43]. The criterion to be minimised over the cross-validation set is the Energy score. This score comprises a multivariate generalisation of the more common Continuous Ranked Probability Score (CRPS). It is a proper skill score already employed by [24] for the evaluation of density forecasts of (u, v)-wind. It will also be considered as a lead score for the out-of-sample evaluation of the skill of the calibrated ensemble forecasts.

Based on this cross-validation exercise, it was found that an optimal value for the forgetting factor would be $\lambda = 0.996$ which corresponds to an equivalent number of observations $n\lambda = 250$ (or in other words 125 days). Additional details and explanations can be found in [5].

5. Assessment of CPS reliability in wind speed mode

The improvement in skill described in [5] seems to originate from the complete calibration of the ensemble forecasts, through translation and dilation. Consequently the main aim at this point is verifying the bivariate reliability of the ensemble forecast of (u, v)-wind before and after calibration.

Reliability diagrams [46] are graphs of the observed relative frequency of an event plotted against the forecast probability of an event. This effectively tells the user how often (as a percentage) a forecast probability actually occurred. In theory, a perfect forecast system would result in forecasts with a probability of p being consistent with the observed frequency. Hence when plotting a reliability diagram comparisons are made against the diagonal. The technique for constructing the reliability diagram is similar to that for calculating the ROC (Relative Operation Characteristics) score, but instead of plotting the hit rate against the false alarm rate, the hit rate is calculated only from the sets of forecasts for each probability separately. It is then plotted against the corresponding forecast probabilities. This work is focusing mainly on extreme winds following closely Deliverable 5.5 (Early warnings and alerts of extreme wind events utilising DWD objective weather type classification methodology and ECMWF EPS Extreme Forecast Index [47]). For consistency with Deliverable 5.5 [48] most of the results in this section are valid for the greater area of Germany as defined in Deliverable 5.5 [48].

- *Reliability assessment over Germany over a two year period (2008 & 2009)*

Reliability diagrams have been constructed for a set of thresholds to study the impact of calibration for different wind speed categories. Threshold values used in this assessment are as follows:

Normal mode categories (wind speed higher than)

- 5 m/s (Appendix A: reliability diagrams T+12, T+24, T+48, T+72, T+96 & T+120)
- 10 m/s (Appendix B: reliability diagrams T+12, T+24, T+48, T+72, T+96 & T+120)

Extreme mode categories (wind speed higher than)

- 15 m/s (Appendix C: reliability diagrams T+12, T+24, T+48, T+72, T+96 & T+120)
- 17.5 m/s (Appendix D: reliability diagrams T+12, T+24, T+48, T+72, T+96 & T+120)

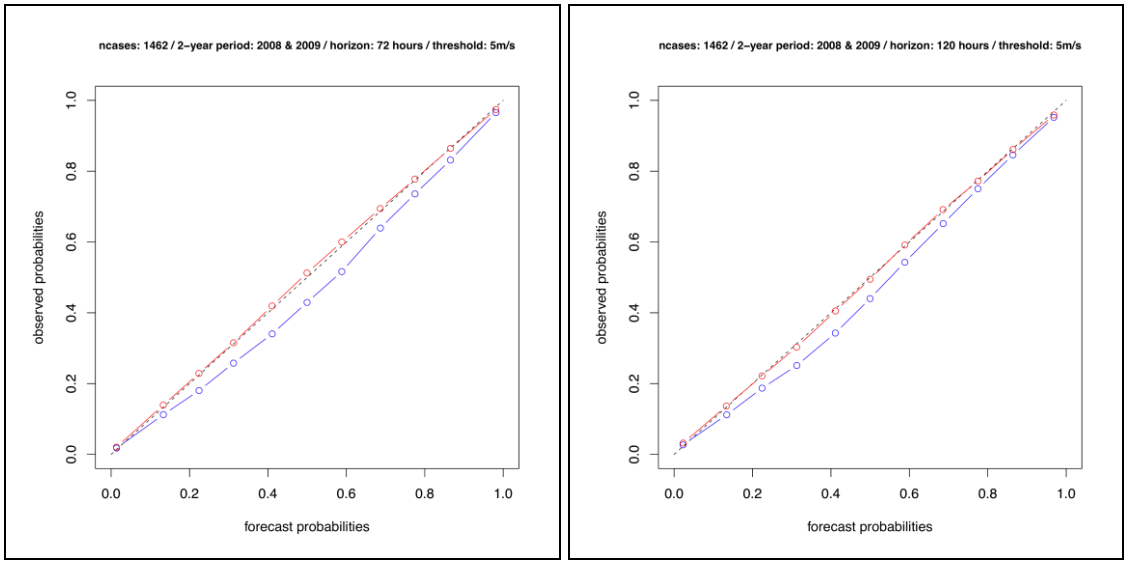


Figure 5.1: *T+72 (day 3) & T+120 (day 5) reliability diagrams for >5 m/s (blue lines correspond to raw ensembles, while red lines to calibrated ensembles)*

Studying closely the full set of reliability diagrams in Appendix A, it becomes clear that calibration results in almost perfect reliability, with red lines (corresponding to calibrated ensembles) being close to the diagonal. Data used here belong to eight seasons (DJF – MAM – JJA – SON 2008 & DJF –

MAM – JJA – SON 2009), while both the 00 and 12 UTC forecasts have been used. Figure 5.1 contains two snapshots of inter-comparison between raw and calibrated ensembles for the category of >5 m/s wind speed. It is obvious that almost perfect reliability scores for the T+72 and T+120 horizons have been achieved for calibrated ensembles.

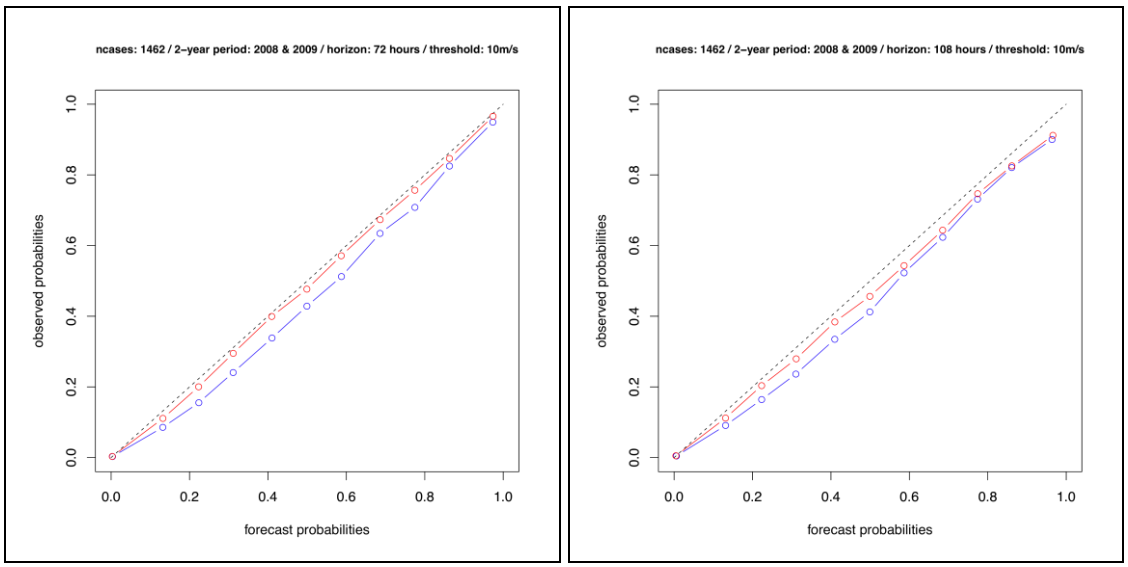


Figure 5.2: *T+72 (day 3) & T+120 (day 5) reliability diagrams for >10 m/s (blue lines correspond to raw ensembles, while red lines to calibrated ensembles)*

Figure 5.2 contains snapshots of raw and calibrated ensembles for the category of >10 m/s. Almost perfect reliability (slightly over-forecasting) is obvious for T+72 forecast horizon. Substantial improvement for calibrated ensembles over raw ones was also found for the remaining forecast horizons between T+12 and T+120 (more details can be found in Appendix B).

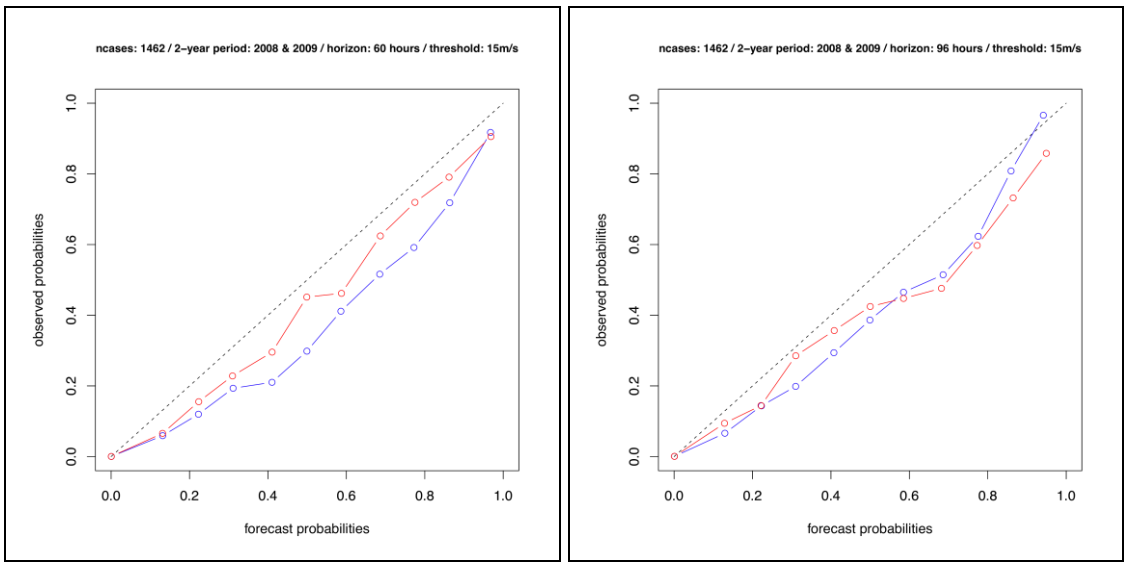


Figure 5.3: *T+60 (day 2.5) & T+96 (day 4) reliability diagrams for >15 m/s (blue lines correspond to raw ensembles, while red lines to calibrated ensembles)*

Figure 5.3 shows snapshots of raw and calibrated ensembles for the (extreme) category of >15 m/s. The T+60 horizon represents the maximum forecast range during which calibrated ensembles are found to be more reliable than raw ensembles for this extreme category (>15 m/s). For horizons longer than T+60 there is no clear signal of some superiority of calibrated ensembles (Figure 5.3, right). More details can be found in Appendix C.

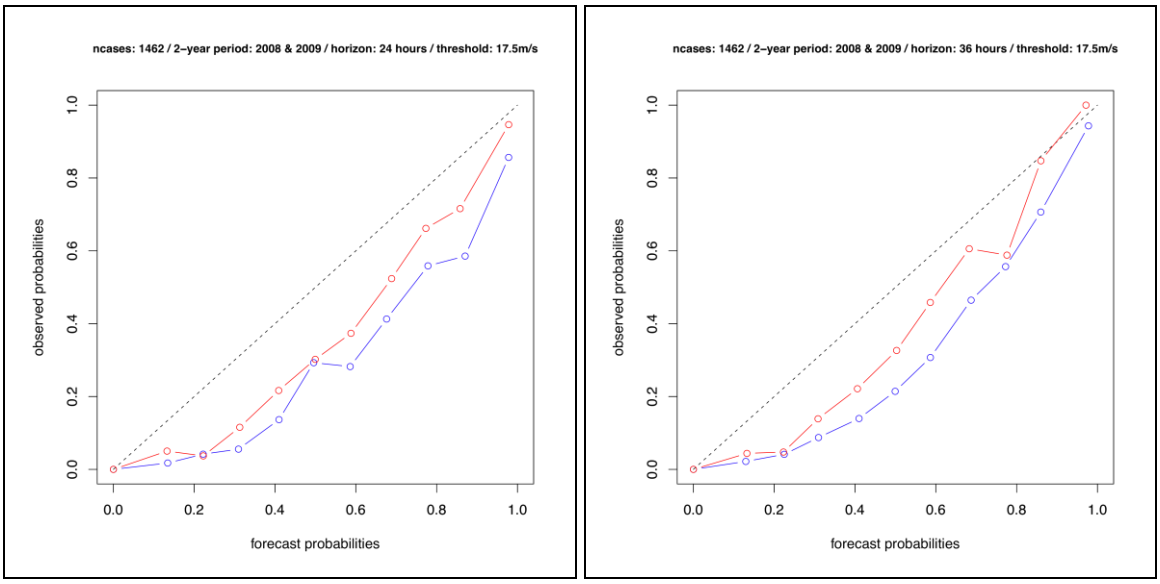


Figure 5.4: *T+24 (day 1) & T+36 (day 1.5) reliability diagrams for >17.5 m/s (blue lines correspond to raw ensembles, while red lines to calibrated ensembles)*

Figure 5.4 illustrates snapshots of raw and calibrated ensembles for the (extreme) category of >17.5 m/s. The superiority of calibrated ensembles capable of producing more reliable forecast guidance in cases of extreme events (>17.5 m/s) is obvious in both schemes. For horizons between T+48 and T+120 there is no clear signal of such superiority of calibrated ensembles over the raw ones (more details can be found in Appendix D).

Note 1: The selected value of 17.5 m/s corresponds to a possible cut off value (25 m/s) at a typical hub height (100 meters). This value has been estimated by utilising the power law vertical wind profile and the assumption of neutral stability conditions.

- *Reliability assessment for Germany for different seasons (2008 & 2009)*

Investigating the variation of skill over different seasons (i.e., the existence of a seasonality concerning the results so far), reliability diagrams were constructed for different seasons and various thresholds. Results for the >10 m/s category (threshold) are presented below. More details can be found in Appendix E and Appendix F.

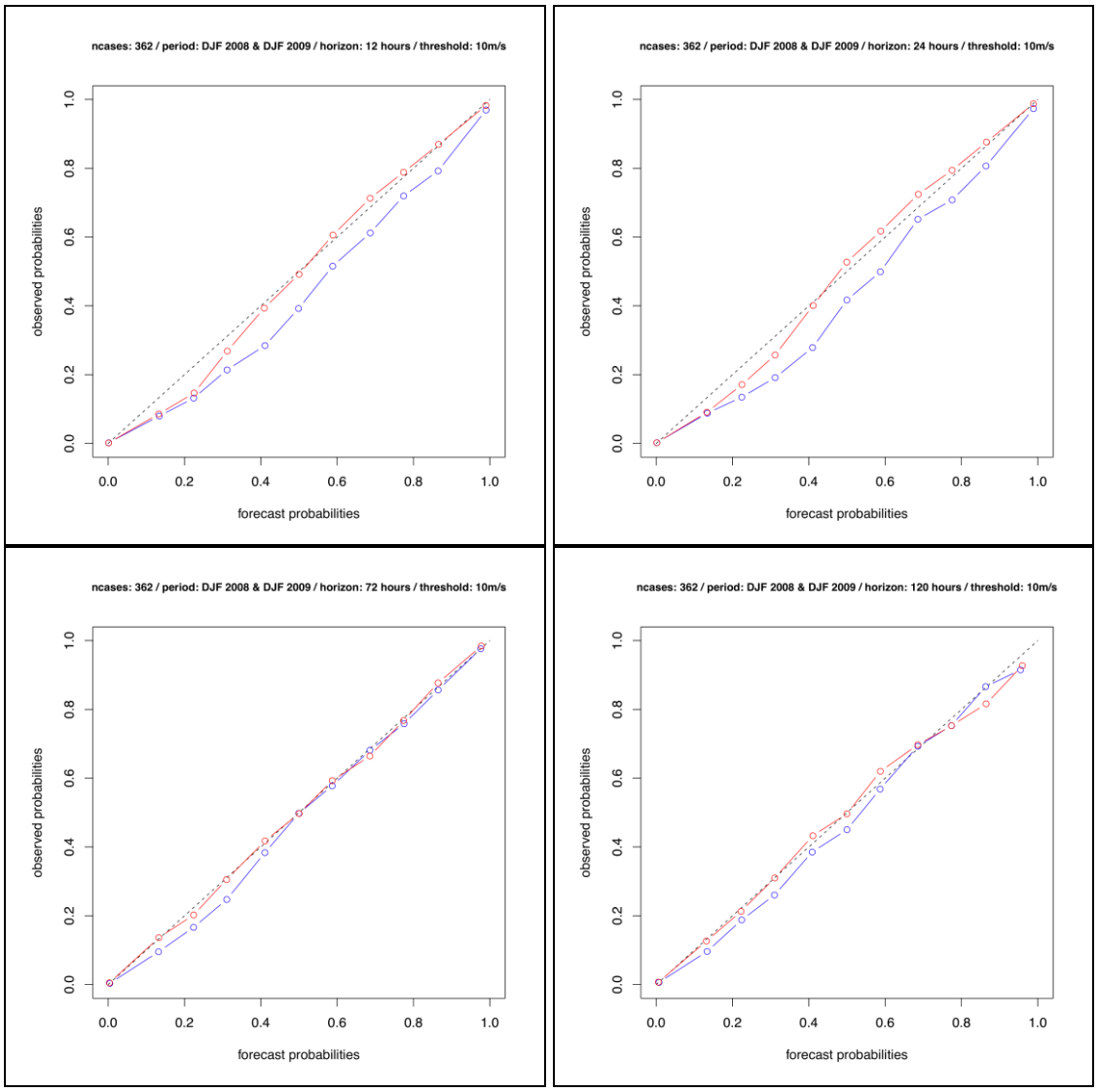


Figure 5.5: *T+12 (top left), T+24 (top right), T+72 (bottom left) & T+120 (bottom right) reliability diagrams for category >10 m/s for 2 winters (2008 & 2009)*

Figure 5.5 contains T+12, T+24, T+72 and T+120 reliability diagrams for the category of >10 m/s over a period of two winters (2008 & 2009). Both the 00 and 12 UTC forecasts have been used comprising a total of 362 cases. The raw ensemble exhibit a positive bias for very short (T+12, T+24) lead times as forecasted probabilities are always higher than the observed probability of the event (>10 m/s). The calibration corrects this deficit effectively. For T+72 and T+120 the uncalibrated ensemble shows a good reliability. Calibrated ensembles manage to achieve perfect reliability. Summarising results are given in Table 5.1.

Table 5.1: *Results of skill evaluation for Germany for different seasons*

horizon in hours	main characteristics for category: >10 m/s
T+12 typical very short-range	substantial improvement after calibration for all seasons highest scores (best reliability results) for autumn
T+24 typical short-range	substantial improvement for all seasons highest scores for spring
T+72 beginning of early medium-range	considerable improvement for all seasons but autumn calibration for autumn results to worse forecast guidance
T+120 end of early medium-range	considerable improvement for spring and summer marginal improvement for winter and autumn

The findings of Table 5.1 are in agreement with results already found and presented in previous subsections. The same investigation was performed for the extreme category of >15 m/s but the number of cases falling in this category was too small to lead to some robust results. That is why a larger area – such as the European one – had to be utilised (as shown in the next Section).

- *Reliability evaluation for Europe for different seasons (2008 & 2009)*

An effort to estimate the impact of calibration over the whole European area is made by utilising reliability diagrams for different seasons putting emphasis on the extreme category of >15 m/s. More details can be found in the corresponding Appendix G and Appendix H.

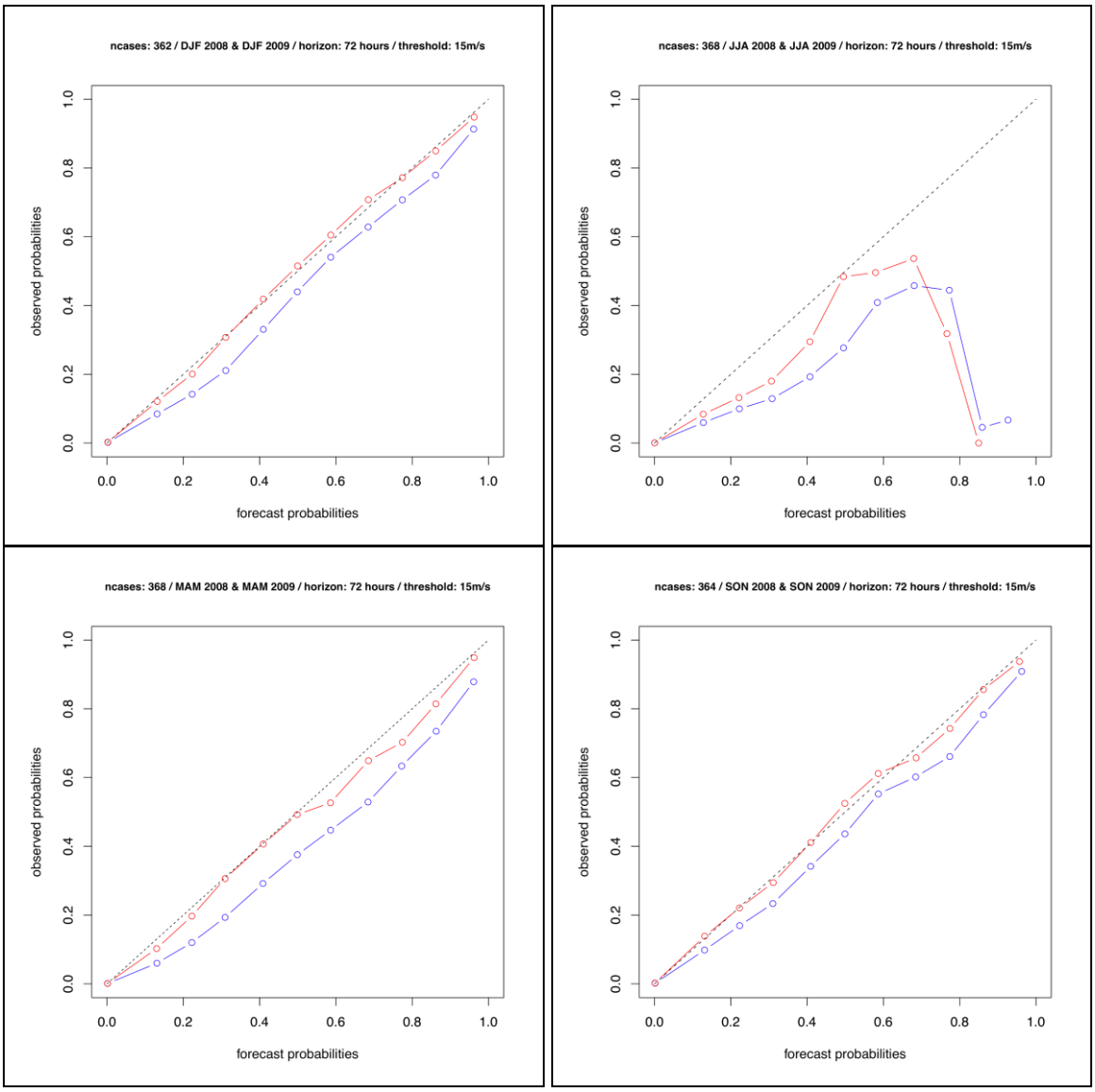


Figure 5.6: *T+72 reliability diagrams for >15 m/s category for 2 winters (top left), 2 summers (top right), 2 springs (bottom left) and 2 autumns (bottom right).*

Figure 5.6 shows the reliability assessment divided by season for the T+72 forecast horizon. The superiority of calibrated ensembles is evident for the winter period achieving almost perfect reliability. In the summer season the lack of high wind speed events deteriorates the reliability for the raw and calibrated ensemble. The superiority of calibrated ensembles is evident for spring and autumn.

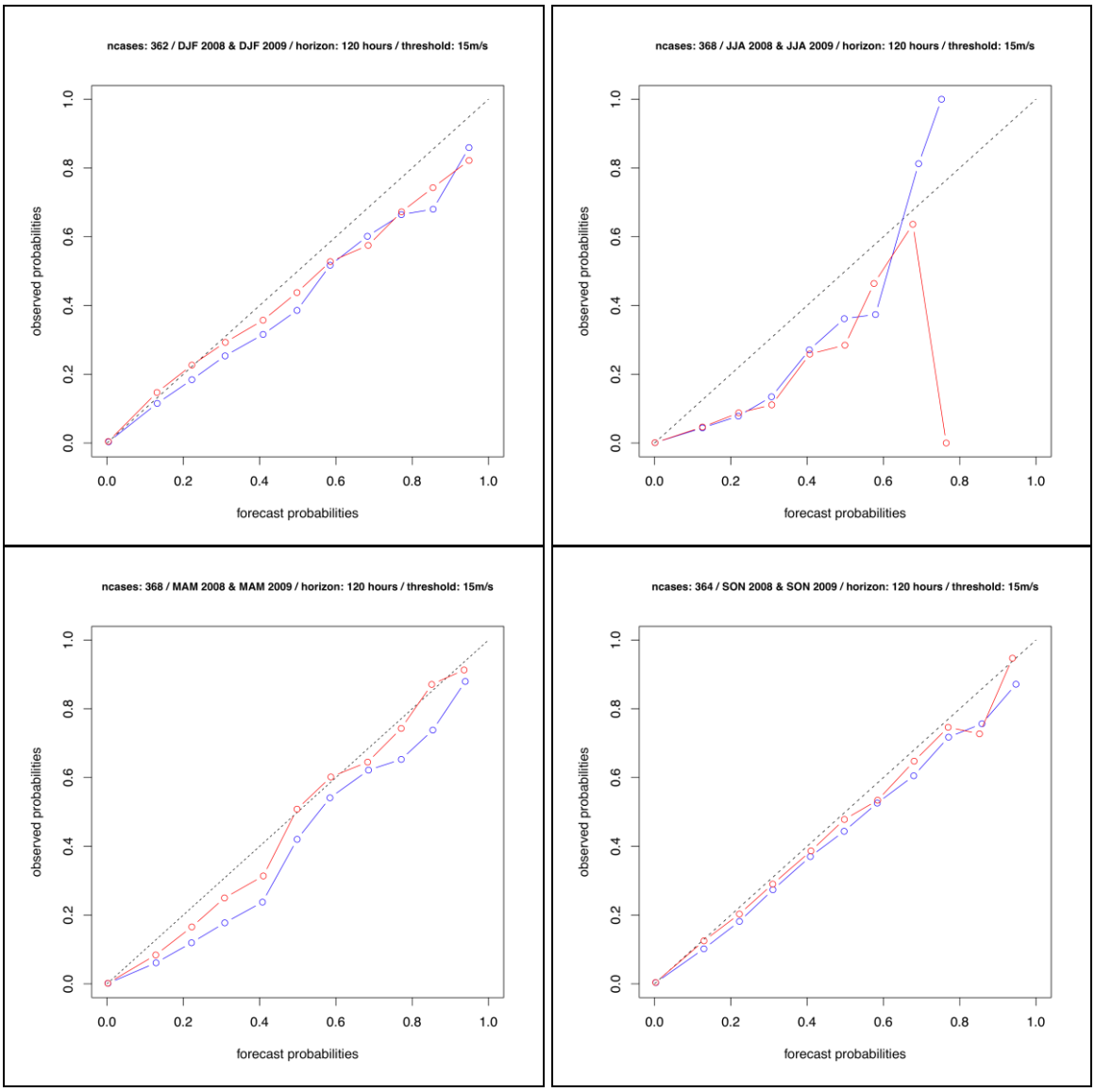


Figure 5.7: *T+120 reliability diagrams for >15 m/s category for 2 winters (top left), 2 summers (top right), 2 springs (bottom left) and 2 autumns (bottom right).*

Figure 5.7 shows the reliability assessment divided by season for the T+120 forecast horizon. The reliability of the raw ensemble is very similar to T+72, except the summer season that is affected by small sample sizes. The calibration improves the reliability considerably, however, it must be stated that the perfect reliability (as for T+72) is not achieved.

6. Assessment of Talagrand diagrams in wind speed mode

Further evaluation of the comparative skill and reliability of the ensemble forecasts of (u, v)-wind before and after calibration, has been performed by utilising the so-called Talagrand diagrams (rank histogram). The rank histogram permits a quick examination of some qualities of the ensemble [49]. Consistent biases in the ensemble forecast will show up as a sloped rank histogram; a lack of variability in the ensemble will show up as a U-shaped, or concave, population of the ranks. Furthermore, the rank histogram may be useful for more than just evaluating the forecast quality. Hamill and Colucci [50 & 51] and Eckel and Walters [52] also show how rank histograms provide information that may be used to recalibrate ensemble forecasts with systematic errors, thus achieving improved probabilistic forecasts.

While it is common for operational centres to produce probabilistic forecasts from their ensembles as if the ensembles were random samples from the same distribution as the truth, in fact many operational centers construct their ensembles under different assumptions. In our case, the singular vector method [38] used at the ECMWF generates initial perturbations that project strongly on the forecast modes where errors grow most quickly. This constitutes a sort of nonrandom sample, where the extremes of the forecast probability density function may be sampled more frequently than the centre of the distribution. The interpretation of rank histograms under such different sampling strategies is not clear. Nevertheless, a wide range of Talagrand diagrams are utilised during our investigation, while emphasis is given on extreme events.

- *Skill assessment for Europe utilising Talagrand histogram*

As already mentioned, although calibrated data sets have been prepared for a period of three years (December 2006 to November 2009), evaluation of the skill improvement has been performed over a period of the two last years, since the first year of data (DJF – MAM – JJA – SON 2007) has been employed as the training set [5].

Bin diagrams shown below are estimated over a period of eight seasons (i.e., DJF – MAM – JJA – SON 2008 & DJF – MAM – JJA – SON 2009) for Europe, while all 00 and 12 UTC available forecasts are being considered. Figure 6.1 contains the T+12 frequencies of the 52 Talagrand bins for both raw (left panel) and calibrated ensembles (right panel). More details can be found in Appendix I containing the full range of Talagrand bins for raw probabilities. The forecast steps used are: T+12, T+24, T+36, T+48, T+60, T+72, T+84, T+96 & T+120. Same wise Appendix K provides the same information as Appendix I but for the calibrated ensembles.

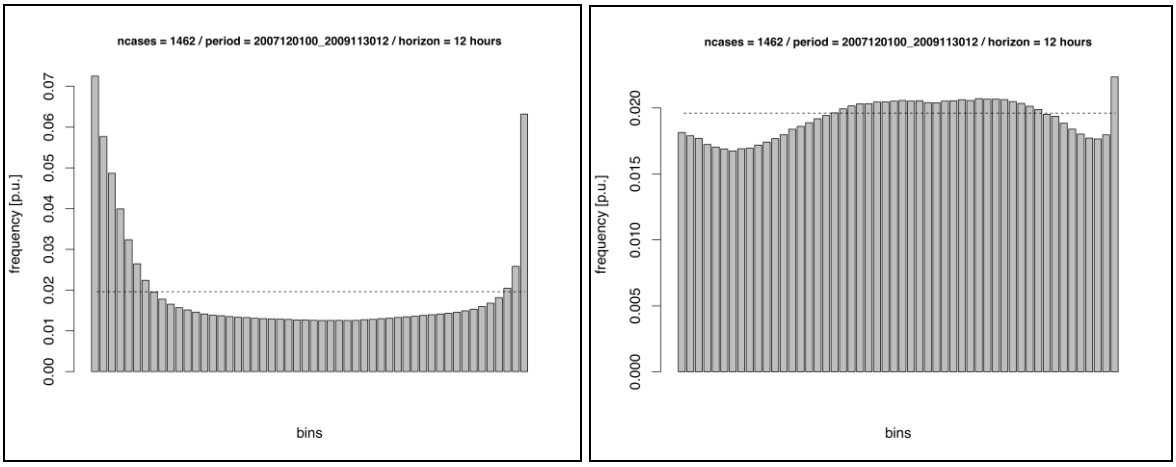


Figure 6.1: *T+12 Talagrand histogram bins for raw (left panel) and calibrated (right panel) ensembles for Europe*

The U-shape dominated raw ensembles of Figure 6.1 (left panel) are a result of higher than expected frequency for observations at the extremes and lower than expected in the middle range of forecasts.

This is an indication of the spread of the EPS being too small for the range of T+12 hour horizon. Calibration seems capable of correcting this deficiency resulting in a more flat distribution, although its final shape is slightly different than an anticipated flat one. This means that calibration is not perfect but at least manages to reduce substantially both the outer left and right bin populations. The one on the right is of great importance since it is linked to possible missed extreme wind cases.

Nevertheless, studying closer Figure 6.1 it seems that calibrated ensembles have a higher chance of capturing events falling into the two tails of observed wind speed distribution during the very short-range.

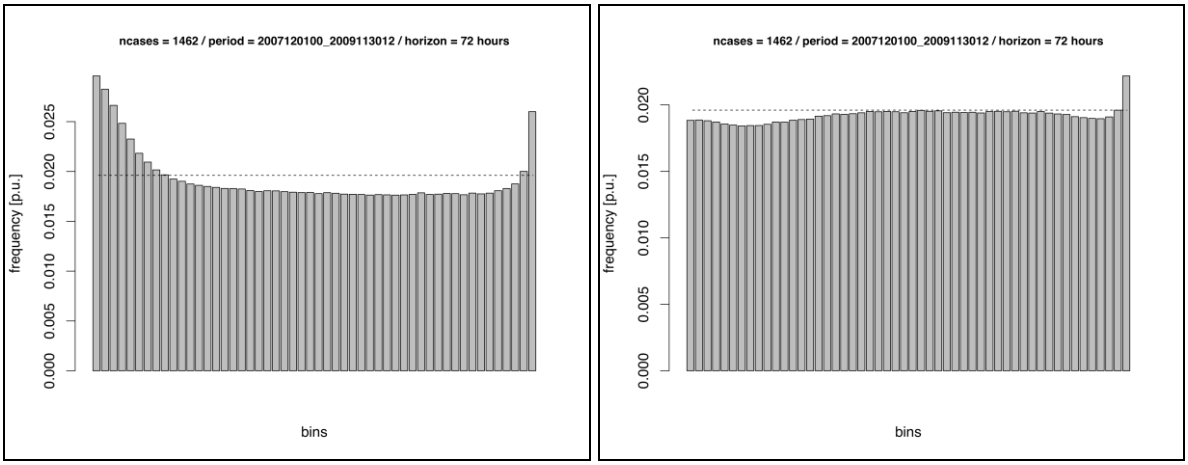


Figure 6.2: *T+72 Talagrand bins for raw (left panel) and calibrated (right panel) ensemble for Europe*

The U-shape (quite pronounced for the T+12 horizon) remains and characterizes the shape of T+72 raw ensembles (Figure 6.2 – left panel), suggesting an “under-dispersiveness” of the EPS spread, while calibration removes most of this deficiency (Figure 6.2 – right panel). The calibrated left outer bin becomes slightly smaller than an anticipated optimal one, while the outer right one is slightly larger than an optimal one (the one corresponding to a constant value).

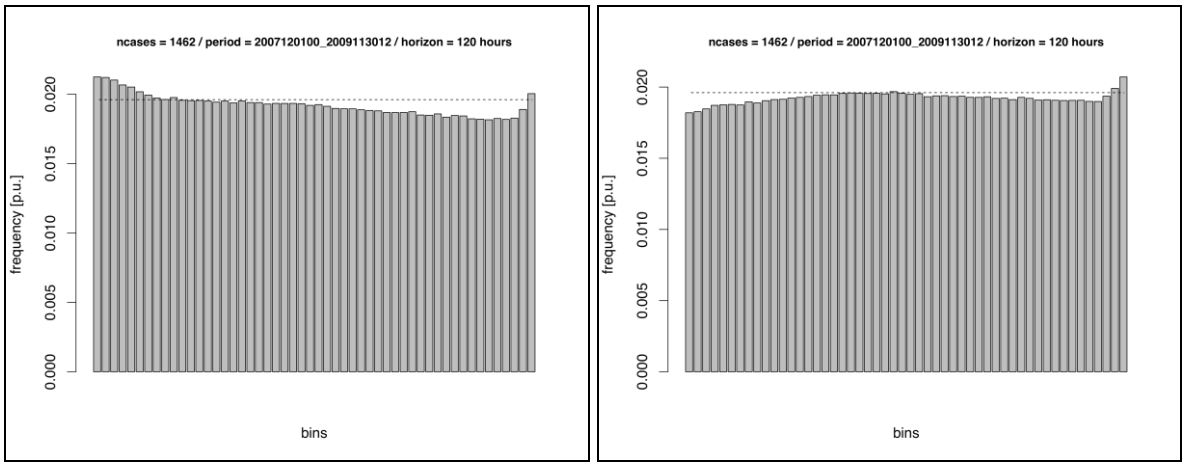


Figure 6.3: *T+120 Talagrand histogram for raw (left panel) and calibrated (right panel) ensembles for Europe*

By day 5, raw ensembles appear to have a profile resembling to the anticipated optimal one (flat) although the outer bins appear to be slightly higher than an optimal one (Figure 6.3 – left panel). This profile has a slight slope to the right which suggest that the raw ensembles overforecast winds.

A similar profile but with an opposite slope is evident for the calibrated ensembles (Figure 6.3 – right panel), although the percentage of the outer right bin is slightly higher than the one corresponding to

raw ensembles, which suggests that calibrated ensembles might miss more wind events (falling in this outer bin category). A detailed description of the percentage of events falling both in the left and right outer bin categories for all forecast steps considered is given in Figure 6.4.

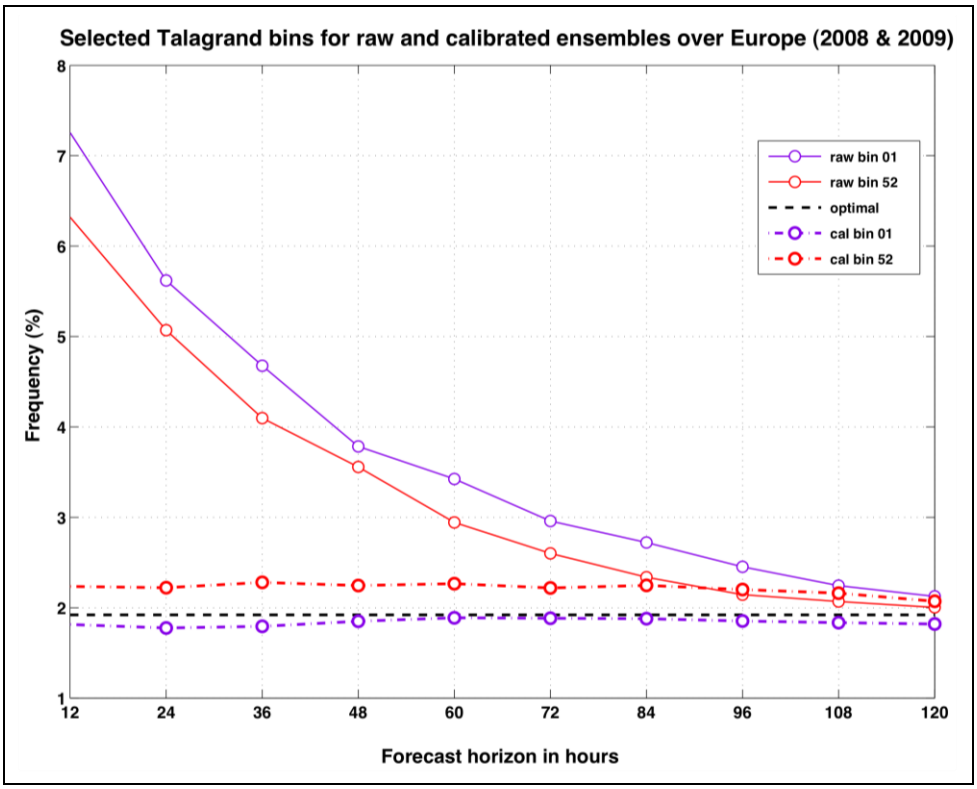


Figure 6.4: *Talagrand outer left (blue) and right (red) bin percentages for raw (solid) and calibrated (dashed) ensembles for Europe for a two-year period (2008 & 2009).*

Concerning the number of the outer left outliers, calibration seems to perform quite well during the entire period of the short- and early medium-range. On the other hand, concerning the events belonging in the outer right bin (raw bin 52), it seems that calibration improves the spread of the ensemble up to T+84 (Figure 6.4). For longer lead times, the raw ensemble captures large events slightly better than the calibrated ensemble (red solid line below red dashed line).

Further investigation concerning the two different base times of forecasts (00 and 12 UTC initialisation time) were carried out as depicted in Figure 6.5 for T+72. Figure 6.5 contains only a limited set of snapshots of graphs of Appendix L (bins for raw ensembles 00 UTC for Europe for 2 years), Appendix M (bins as in Appendix L but for calibrated ensembles 00UTC), Appendix N (bins as in Appendix L but for raw ensembles 12 UTC) and Appendix O (bins as in M but for calibrated ensembles 12 UTC). Calibration performs equally well for both sets of 00 and 12 UTC forecasts, leading to a smaller set of numbers concerning the population of the outer bins. Note, that 00 UTC and 12 UTC have always been calibrated jointly but have been split only for evaluation in the previous investigation.

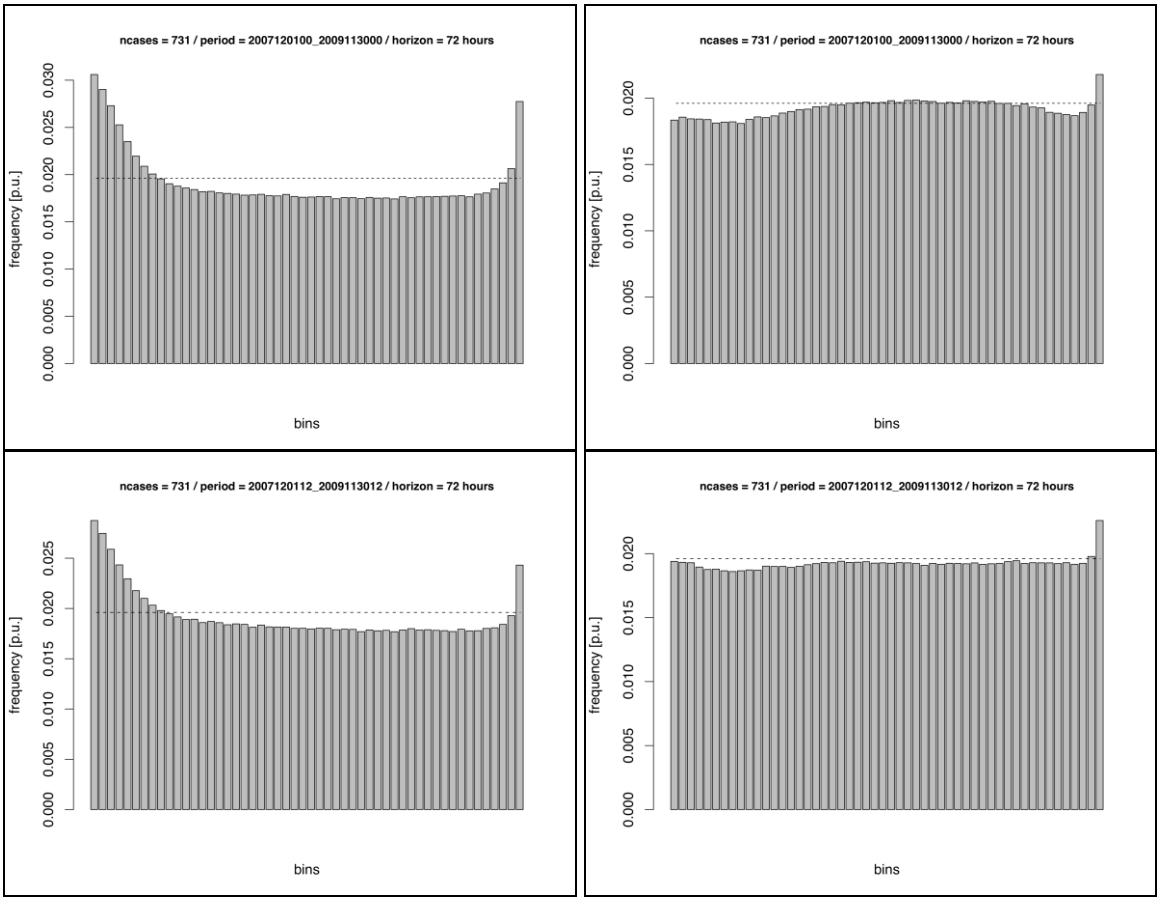


Figure 6.5: *T+72 Talagrand histogram for raw 00 UTC (upper left) and 12 UTC (lower left) against calibrated 00 UTC (upper right) and 12 UTC (lower right) ensembles for Europe*

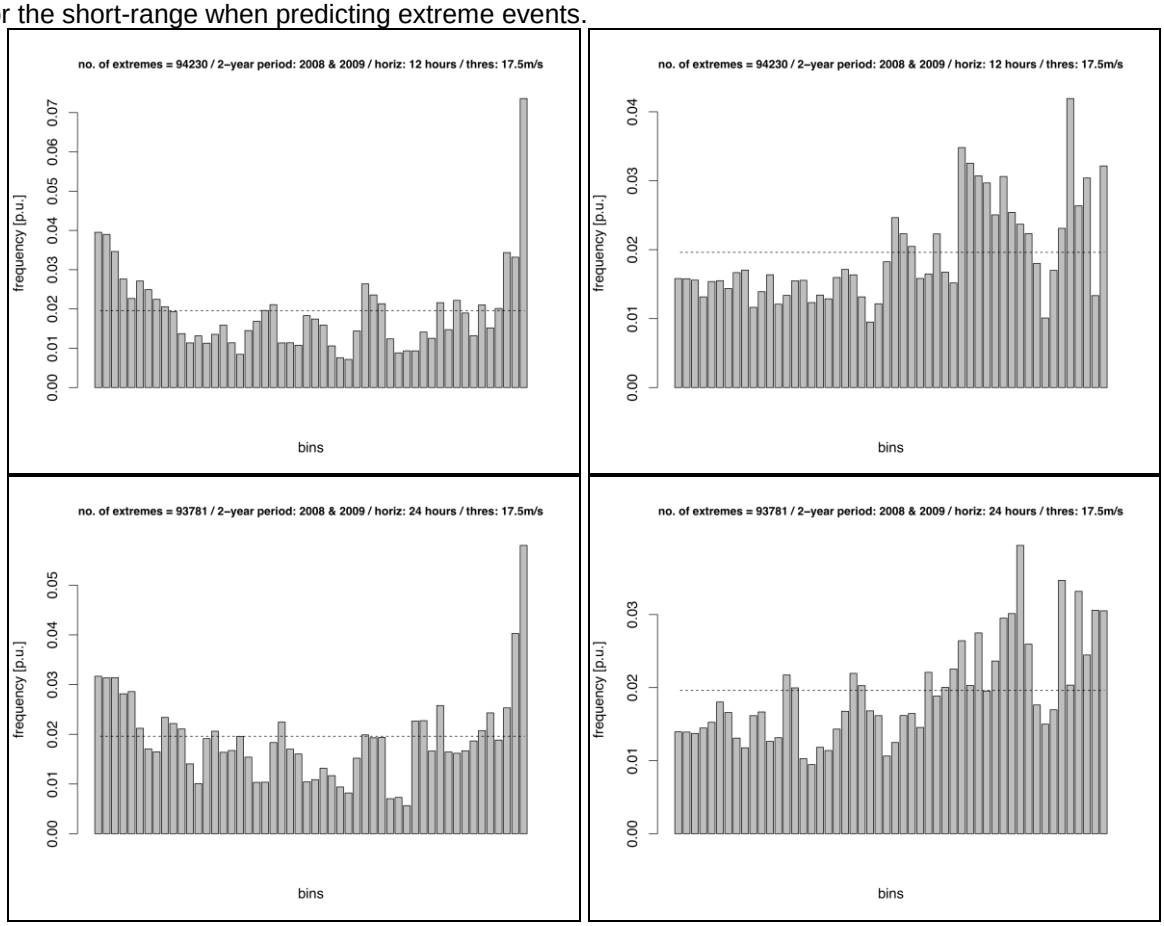
7. Focusing on wind extremes (Talagrand bins for extremes)

Talagrand bin histograms used in the previous Section had been constructed for the full spectrum of wind events. This spectrum includes a certain portion of extreme events that is interesting to end-users in the wind energy business that are addressed by the SafeWind project. Thus, a series of Talagrand diagrams are constructed for certain sets of events exceeding critical wind speed threshold:

- greater than 15 m/s
- greater than 17.5 m/s (corresponding to a 25 m/s cut-off at typical hub height of 100 meters)
- greater than 20 m/s

These three thresholds are considered as extremes categories. The >15 m/s category corresponds to 3.1% of total wind events, while the >17.5 m/s category to 1.4%. The most extreme category (>20 m/s) corresponds to only 0.6% of the total wind events (considered as a rare extreme event) for Europe. As in the previous sections the verification is based on ECMWF 10m analysis winds.

Talagrand diagrams were constructed for raw ensembles and CPS (calibrated) ensembles. Figure 7.1 shows the T+12, T+24, T+72 and T+120 Talagrand diagrams of raw (left panel) and calibrated ensembles (right panel) for events falling in the >17.5 m/s extreme category, over a period of two years (2008 & 2009) for both 00 and 12 UTC forecasts valid for Europe. It is obvious that calibration reduces the population of both outer bins in the short range. The number of left outliers drops below the optimal value of 1.92%, while the number of right outliers (most extreme events) is being reduced substantially. Thus, it becomes obvious that calibration is capable of providing the user with more useful ensembles for the short-range when predicting extreme events.



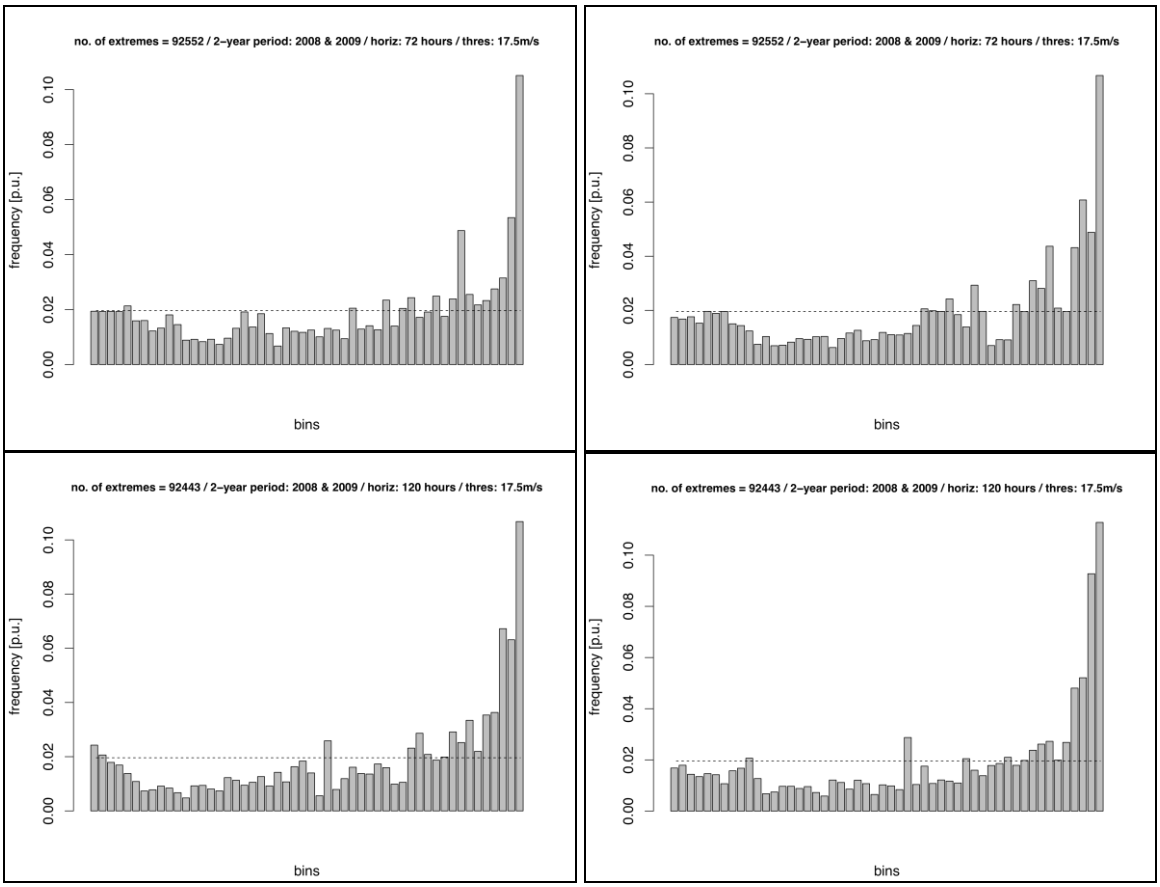


Figure 7.1: *T+12, T+24, T+72 and T+120 (from top to bottom) Talagrand histogram for raw (left) and calibrated (right) ensembles for Europe for extreme wind events >17.5 m/s.*

Unfortunately, CPS loses its beneficial effect of capturing more extremes than the raw ensemble when entering the early medium-range (T+72 hours) and medium-range (T+120 hours). Almost no difference to the Talagrand histogram of the raw ensemble can be noticed.

Figure 7.2 summarizes the findings so far concerning the time evolution of the population of the outer right bin (labelled as bin 52) during the short- and early medium-range. In the lower part of the graph the inter-comparison between raw and calibrated ensembles – valid for the full spectrum of wind events – is given by coloured bars. In the upper part of the graph the blue line corresponds to the percentage of missed extreme events by raw ensembles, while the red line corresponds to the ones missed by calibrated ensembles. The horizontal black line represents the optimal frequency (percentage) in a typical Talagrand bin diagram. More details can be found in Appendix T (raw ensembles) and Appendix U (calibrated ensembles).

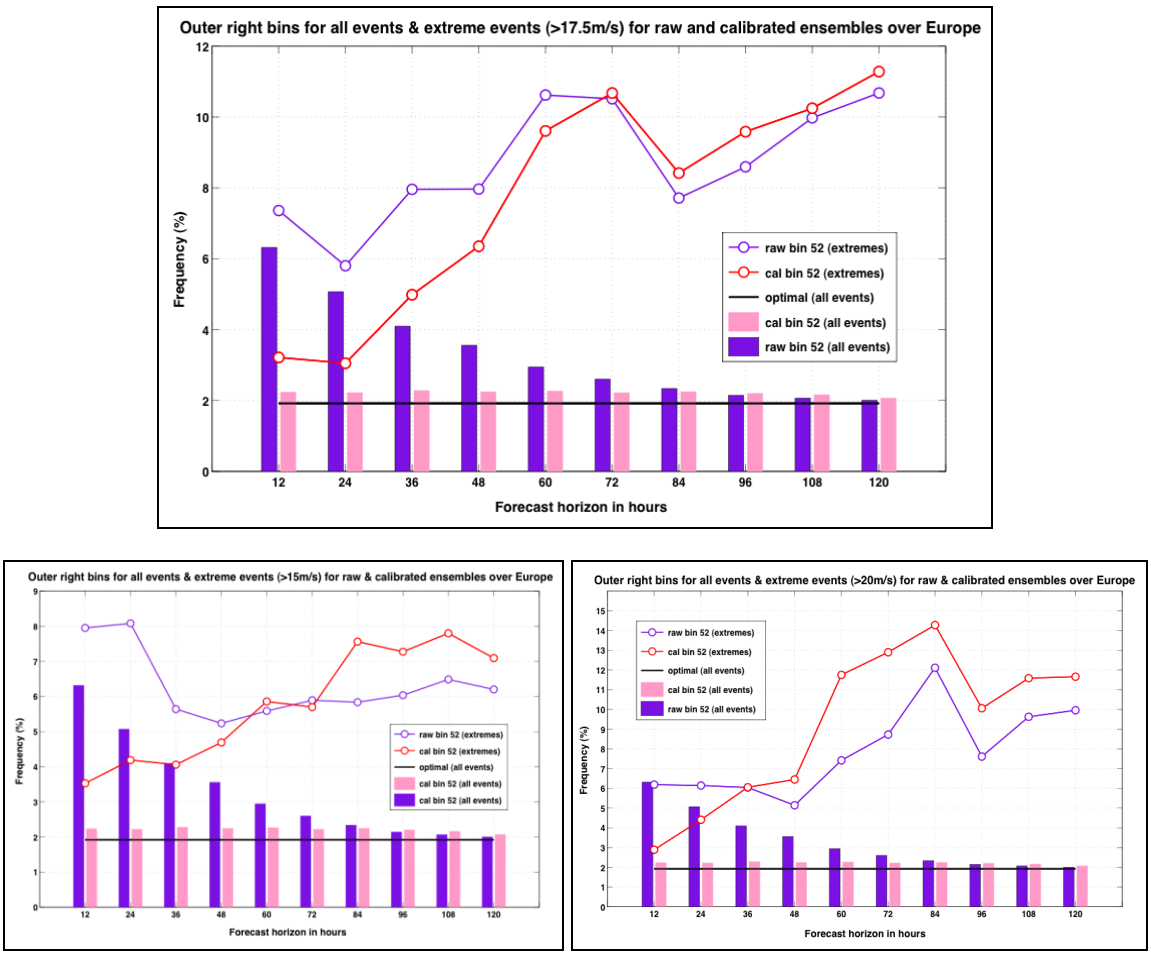


Figure 7.2: *Outer right bins (no. 52) for all events & extreme events >17.5 m/s (top), >15 m/s (bottom left) and >20 m/s (bottom right) for raw and calibrated ensembles for Europe (2008 & 2009).*

It becomes obvious that calibration (concerning extremes) proves to be beneficial till 60 hours. Beyond this forecast horizon the blue (raw ensemble) and red (calibrated ensemble) line crosses each other and the calibrated ensembles is less capable to capture extremes of category > 17.5 m/s compared to the raw ensemble (Figure 7.2, top). This deficiency of the CPS ensemble was found to be less pronounced for the extreme category of >15 m/s (Figure 7.2 – left panel), but more pronounced for the extreme category of 20 m/s (Figure 7.2 – right panel).

Details can be found in Appendix R and Appendix S (concerning >15 m/s category), while Appendix V and Appendix W contain details of raw and calibrated ensembles related to the rare >20 m/s wind speed category.

In addition, Figure 7.3 summarizes the population of the outer left bin during the short- and early medium-range for three wind speed categories. Similar to the findings in Figure 7.1 the good performance of the calibration for the outer left bin can be noticed compared to the poor performance of the raw ensemble in the short-range. Beyond T+72 no calibration need can be noticed as the outer left bin has reached the 1.92 % frequency.

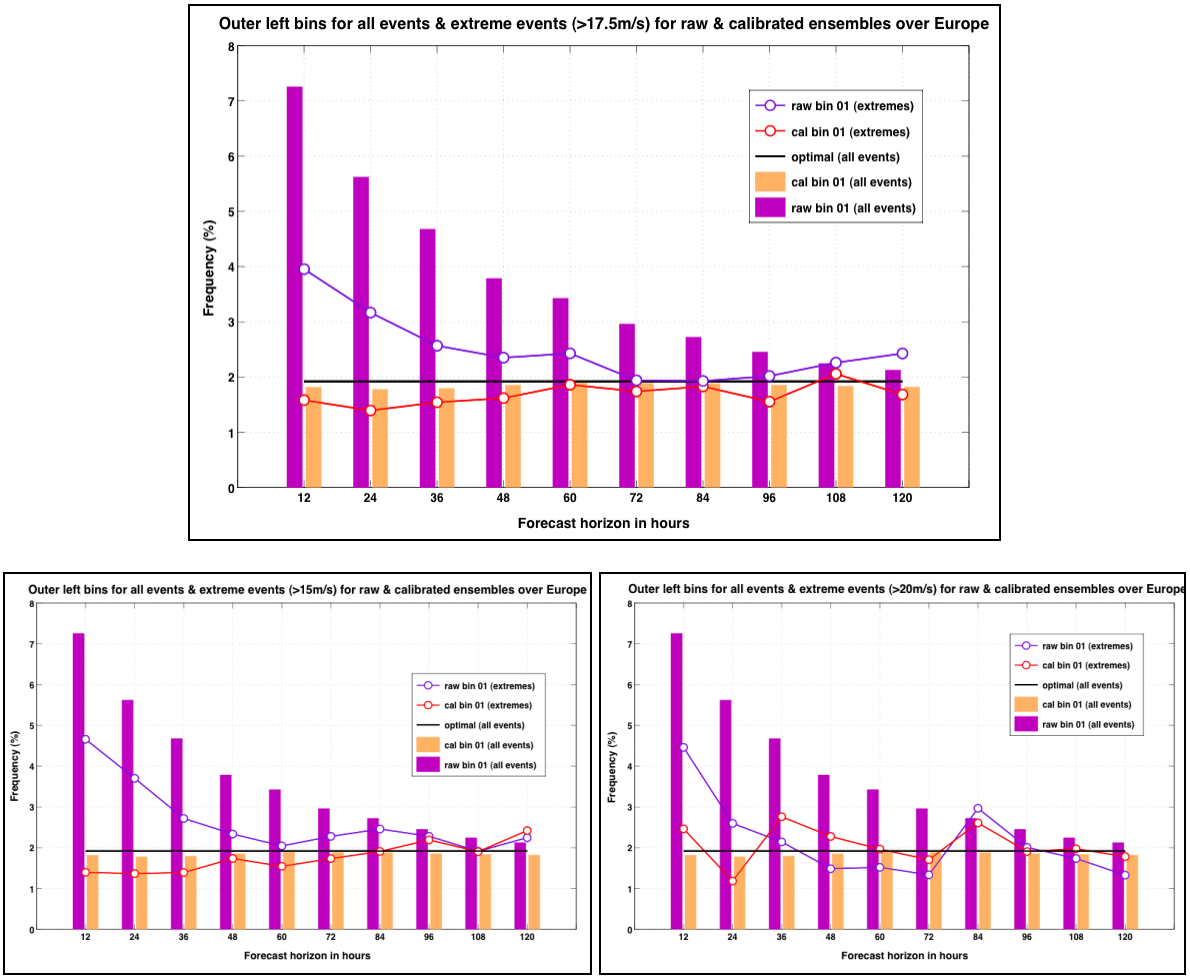


Figure 7.3: *Outer left bins (no. 01) for all events & extreme events >17.5 m/s (top), >15 m/s (bottom left) and > 20m/s (bottom right) for raw and calibrated ensembles for Europe (2008 & 2009).*

8. Skill assessment of CPS in wind power mode

In the previous sections the skill of CPS was assessed in wind speed mode and verification has been done with the ECMWF model analysis. It must be noted that CPS is targeted to match the (smooth) model analysis. The benefits and limits of CPS with respect to real observations in wind power model are analysed in this section for the example of German wind power predictions. The predictions are computed individually for each of the four control zones and finally aggregated.

8.1 Wind power forecast model

The used wind power forecasting model for Germany is intentionally kept very basic. The data base [54] that hold the information about regional distribution and capacity of wind power deployment in Germany is evaluated on a monthly basis. All wind turbines in Germany including the installed capacity, rated power, date of commissioning and geographical information are listed in the data base and are mapped to the model grid points of the CPS. The number of model grid points is 460 for Germany.
However, not all model grid points are contain installed wind power capacities. For each model grid point the following information has been used by the wind power forecast model:

- installed wind power capacity,
- hub height (weighted according to wind turbines allocated to this model grid point),
- lowest surface roughness length z_0 corresponding to the 20 % quantile given in the grid cell based on 7x4.2 km resolution.

The winds are extrapolated with the logarithmic wind profile for neutral conditions to hub height using the supplied surface roughness length. Unfortunately, the thermal stratification of the atmosphere can not be computed from available ECMWF EPS forecasts. Neither temperature forecasts on model levels nor the surface friction velocity are archived to compute atmospheric stability. Consequently, large over (or under-) estimations of forecasted wind speeds in hub height occur in unstable (stable) conditions. As stable conditions often occur during night, wind power is underestimated during night. During the day wind power forecasts are often overestimated because the vertical wind shear is smaller than suggested by the neutral logarithmic wind profile. Wind speed forecasts in hub height are converted into wind power predictions utilizing a regional power curve developed in the TradeWind project [55]. The power curve is shown in Figure 8.1. More details about the wind power forecasting model can be found in [56].

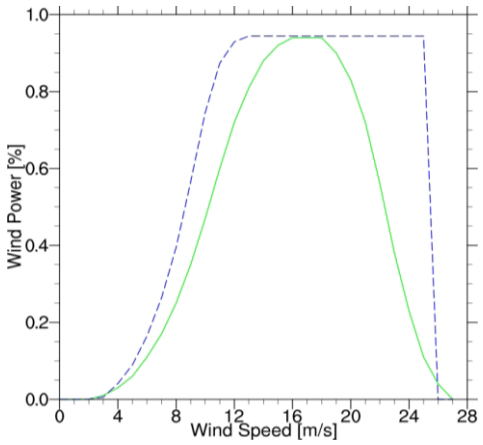


Figure 8.1: *Regional power curve from the TradeWind Project (green) and manufacturer's onshore power curve (blue) for a multi-megawatt wind turbine.*

The usage of 00 and 12 UTC forecasts leads to a twelve hour cycle of the systematic forecast error (not shown). The twelve hour cycle in the root mean square forecast error (RMSE) is caused by this strong systematic forecast error (bias) (Figure 8.2, left).

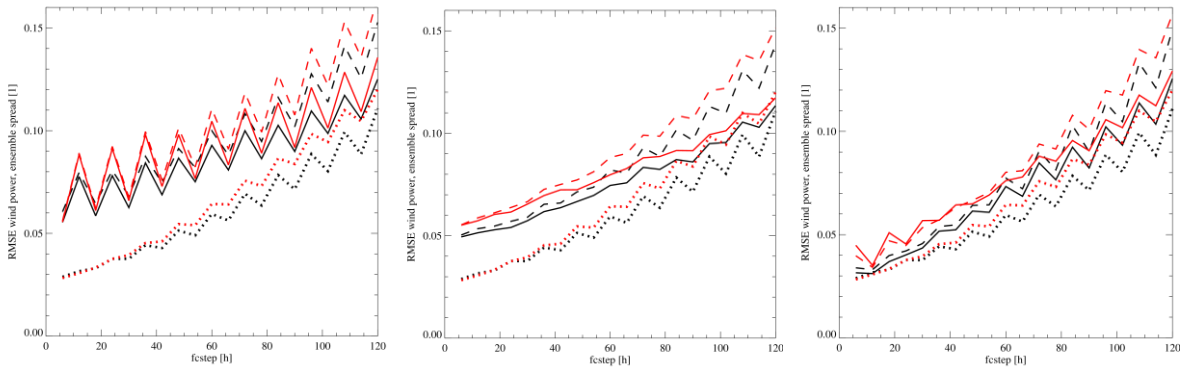


Figure 8.2: *Ensemble wind power forecast spread (dotted line), skill (RMSE) of ensemble mean wind power forecast (full line) and control forecast (dashed line) over forecast step. Wind power forecasts are computed for Germany with raw 10 m winds (red) and calibrated 10 m winds (black). Left: verification against observed wind power and no wind power bias correction. Middle: time-of-the-day dependent wind power bias correction and verification against observed wind power data. Right: verification against simulated wind power (no wind power bias correction).*

The strong diurnal bias in wind power forecasts for Germany when verifying against measured data is not satisfying and a simple bias correction has been applied to the wind power forecast data. The wind power bias correction is dependent on the time of the day and is done for each of the four TSO zones in Germany individually. The bias correction consists only of an additive component, i.e. linear regression is not performed because a linear regression would also affect the ensemble spread. However, the modification of the ensemble spread can be worthwhile to calibrate the ensemble properly, but in that case a more tailored calibration is preferable to correct spread and bias jointly.
When verifying against simulated wind power no wind power bias correction is applied, because forecasted and simulated wind power have the same deficits of not accounting for thermal stratification effects in turbine hub height. It is assumed that the CPS 10 m wind speeds are unbiased with respect to the ECMWF analysis.

8.2 Deterministic forecast verification

The (deterministic) forecast skill, expressed as RMSE normalized with the installed capacity, is shown in Figure 8.2. The wind power bias correction is very efficient to remove the diurnal bias (Figure 8.2, middle). The calibrated (CPS) ensemble mean and the calibrated control forecast (black lines) are substantially better (up to 1 %) than the raw ensemble (red lines). The ensemble mean is already at Day+1 outperforming the (single) control forecast for raw and calibrated forecasts.
The calibration tends to decrease the ensemble spread, defined as root mean square difference of the ensemble members to the ensemble mean, for higher lead-times (dotted line). The black (dotted) line is for CPS and the red line for raw the EPS.
It can be noted that when verified against the analysis (Figure 8.2, right) the wind power RMSE is considerably lowered compared to the verification against measured wind power. Consequently, the spread to (RMSE) skill relation is far better for the verification against analyses. A good spread-skill relation means that ensemble spread and (RMSE) skill are matching each other. In case, the spread is lower (higher) than the skill the ensemble is under (over)-dispersive.

8.3 Probabilistic verification against simulated wind power

Not only wind speed forecast are converted into wind power utilizing the wind power forecast model but also wind analysis data at 10 m height are used to simulate the production of wind power in Germany.

Hence, simulated wind power data has the same deficiencies with respect to thermal stratification effects as the wind power forecasts. Consequently, a diurnal cycle does not occur in the skill of the ensemble mean (Figure 8.2, right).
Thus, simulated wind power data is ideal to evaluate the impact of CPS winds over the raw ensemble disregarding possible deficiencies in the conversion of wind into wind power

Talagrand (Rank) Histograms for all events are shown in Figure 8.3 for forecast Day+1, +3 and +5. The Talagrand Diagram for the raw ensemble (Figure 8.3, left) is skewed to the left indicating that quite often simulated (equivalent to observed) wind power is lower than the lowest forecast members, i.e. the ensemble forecast has a positive bias. This positive bias becomes smaller for higher lead times. The calibrated ensemble leads to an improved Talagrand Diagram (Figure 8.3, right). At Day+1 the calibrated ensemble is slightly overdispersive. For the other lead times the CPS in wind power is able to capture the distribution (including the tails) of wind power events very well. However, the calibrated Talagrand Diagram looks noisy compared to the calibrated Talagrand Diagram of wind speed forecasts for Germany that are given in Appendix Q. It is likely that the cubic of wind speed that is used for wind power forecasts amplifies the differences between the ensemble members substantially.

The CPS wind power forecasts show almost perfect reliability compared to the uncalibrated winds when verifying against simulated wind power (Figure 8.4, left). Calibrated forecast probabilities for the event wind power >50 % of installed capacity match the observed probability quite well. Mostly, the deviation is smaller than the calculated consistency bars. The consistency bars are calculated following a method suggested by [57] in order to take the limited number of cases into account.
The strong positive bias of the uncalibrated ensemble becomes also visible in the reliability diagram, since the red (dashed) line lies completely under the diagonal. This means that for all forecast probability classes the observed frequency of the event (> 50 % of installed wind power) is substantially lower. The same positive bias of the uncalibrated ensemble is observed in the Reliability Diagrams for German wind speed in Figures 5.1-5.4. Those figures also show that the CPS ensemble is very well calibrated or is at least more reliable than the raw ensemble.

The sharpness analysed by the frequency of forecast probabilities (subplot) is very similar for raw and CPS wind power forecasts, i.e. low and high forecast probabilities can be distinguished very well. Concerning the evaluation with respect to extremes it must be stated that already the probability of wind power exceeding 50 % of installed capacity is very low (~8 %) and consequently the occurrence of high forecast probabilities is very low, too.

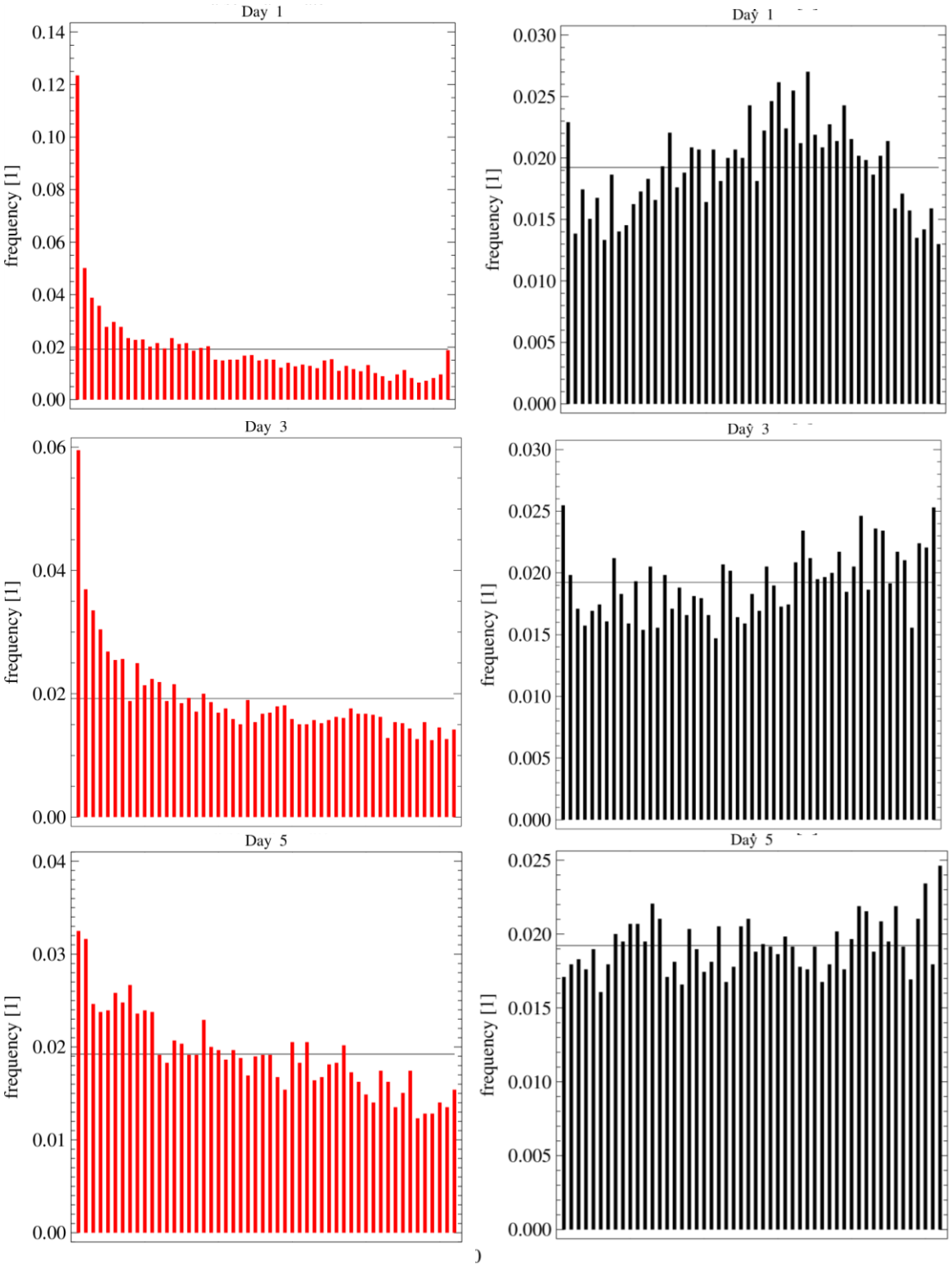


Figure 8.3: *Talagrand Diagram for forecast Day+1, +3 and +5 for uncalibrated (red) and (CPS) calibrated (black) 10 m EPS winds forecasting German wind power. The verification is done against simulated wind power.*

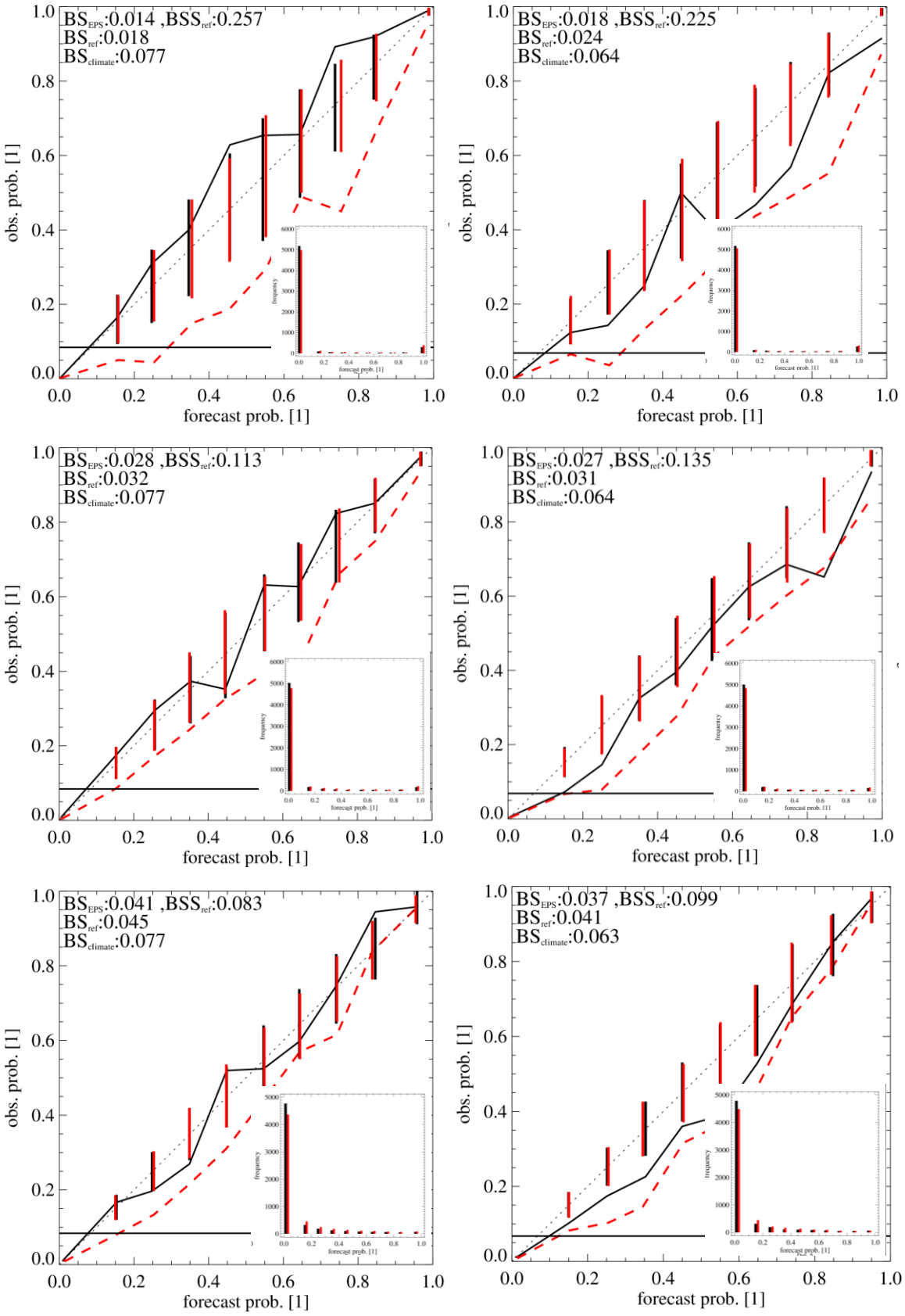


Figure 8.4: *Reliability Diagram for the event wind power >50 % of installed capacity for (CPS) calibrated (black) and uncalibrated (red) 10 m EPS winds forecasting German wind power at forecast Day+1 (top), Day+3 (middle) and Day+5 (bottom). The verification is done against simulated wind power (left) and against real feed-in data including a post-processing for wind power bias correction (right). The vertical bars are consistency bars (90 % confidence) to consider sampling errors.*

8.4 Probabilistic verification with observed wind power

Real wind power production is available on the websites of the four German TSOs in 15 minute resolution. Even though wind power production data has been averaged to hourly values, the spread of both (the raw and the calibrated) ensemble forecast is too small to capture low and high wind power events properly. The Talagrand Diagrams without wind power bias correction at D+3 (Figure 8.5, top) are U-shaped and in contrast to the verification against simulated wind power, no advantage of the

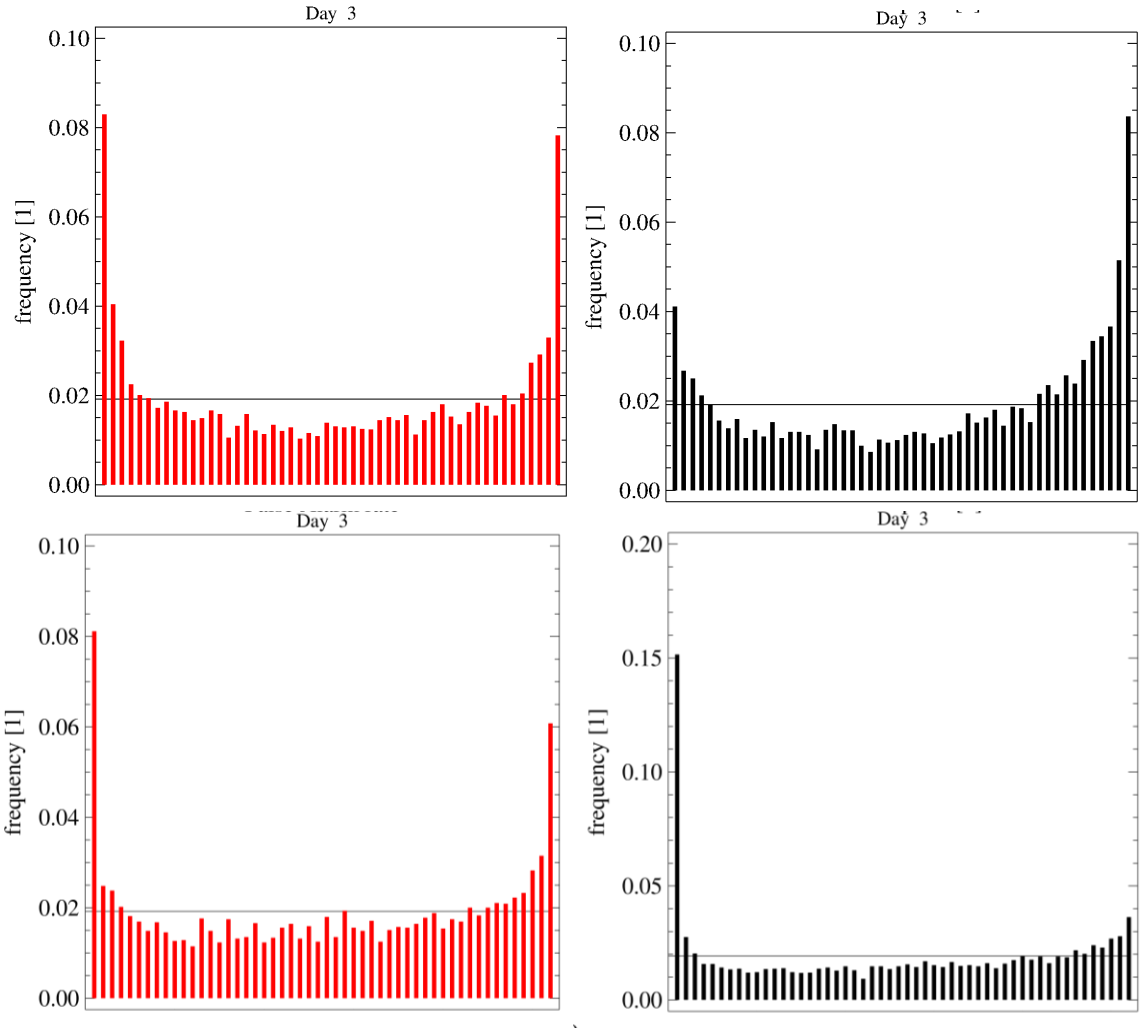


Figure 8.5: *Talagrand Diagram at forecast Day+3 for uncalibrated (red) and calibrated (black) 10 m EPS winds forecasting German wind power. The verification of not biased-corrected (top) and bias-corrected (bottom) wind power forecast is done against observed wind power.*

CPS winds can be noted. The post-processing to remove the large wind power bias does not change the Talagrand Diagram of the raw ensemble (Figure 8.5, left), but the Talagrand Diagram for CPS degrades even further (Figure 8.5, lower right), i.e. low observed wind power values are very often outside the ensemble range. The insufficient spread of the CPS when verified against observations can also be seen in the higher discrepancy of spread and RMSE at Day+3 in Figure 8.2 (middle) compared to Figure 8.2 (right) where the verification is done against simulated wind power.

The lack of spread is even much worse at Day+1 (Figure 8.6) and CPS further increases the inability to capture low wind power values (Figure 8.6, top right). The wind power bias correction degrades the uncalibrated Talagrand Diagram (Figure 8.6, bottom left) while the Talagrand Diagram for CPS winds is improved by the wind power bias correction but still of poor skill (Figure 8.6, bottom right).

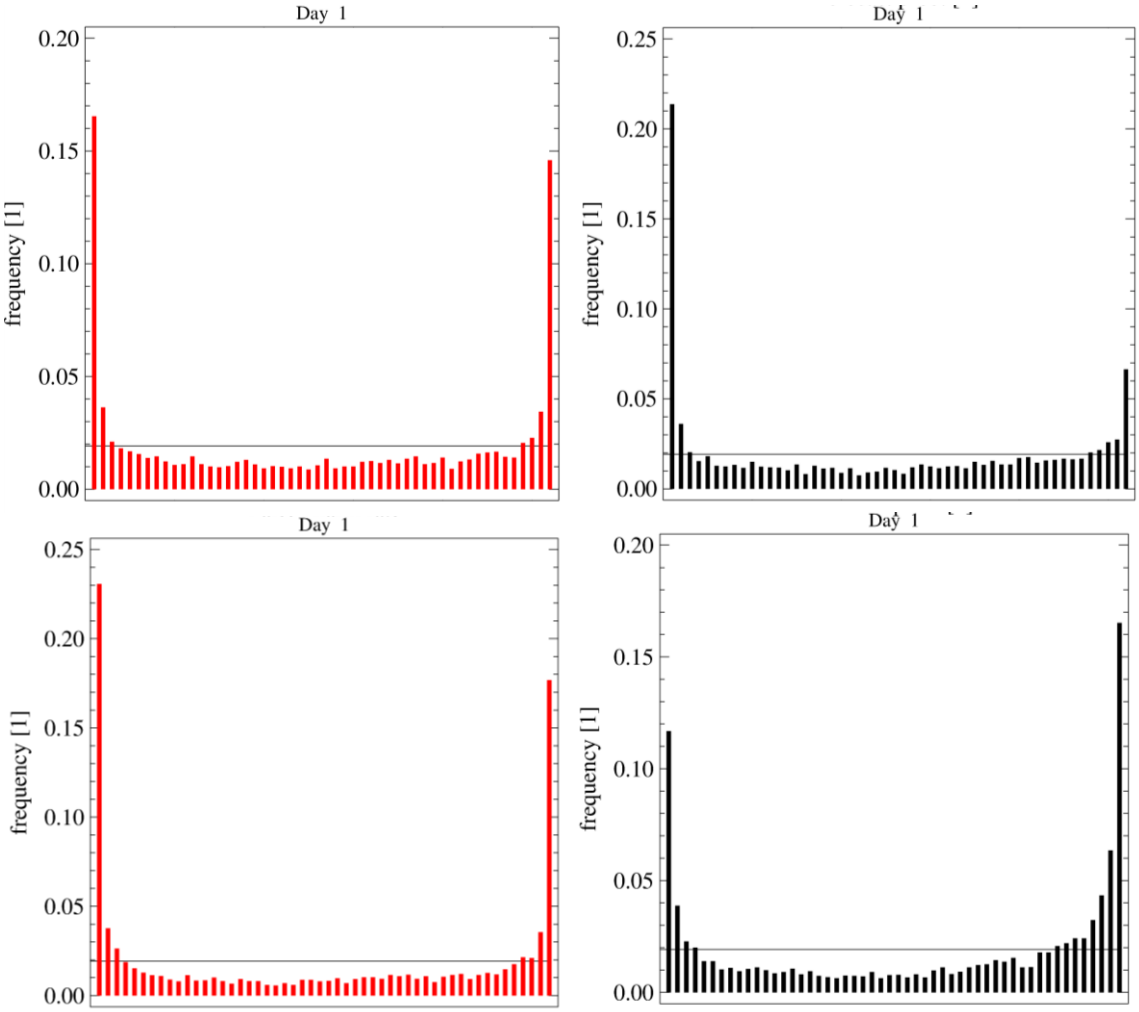


Figure 8.6: *As Figure 8.5, but at forecast Day+1.*

The reliability of the calibrated ensemble predicting wind power >50 % of installed wind power is clearly improved compared to the raw ensemble (Figure 8.4, right). The Brier Skill Score (BSS) printed in Figure 8.4 shows that for short lead times the improvement due to calibration is higher for verification against simulated wind power compared to verification against observed wind power. For higher lead times the improvement through CPS is higher for verification against observations.
However, at Day+5 the CPS ensemble has still a strong positive bias, since the reliability curve in Figure 8.4 (lower right) is still clearly below the diagonal and outside the consistency bars.

8.5 Scoring of the CPS improvement

When assessing the overall skill of a new ensemble system all three important characteristics of a probabilistic forecast must be considered, i.e. reliability, resolution and sharpness. A score that combines all is the Continuous Ranked Probability Score (CRPS). The CRPS is often expressed with respect to a reference system and called Continuous Ranked Probability Skill Score (CRPSS). CRPS or CRPSS is not bounded to certain thresholds that shall be exceeded or not exceeded but samples the whole range of events. Details are given in [58].

In the reference ensemble system uncalibrated 10 m winds are used. Figure 8.7 shows the CRPSS for two control zones (Tennet and 50Hertz) and entire Germany. The verification is done with simulated wind power (from 10 m winds) and also with observed wind power. The CRPSS has a very strong 12 h cycle (double diurnal cycle) that is also observed in the RMSE of the ensemble mean without bias correction (Figure 8.2, middle) This strong (double) diurnal cycle was also diagnosed in wind speed when the CPS was developed and verified utilising the Energy Score [1].

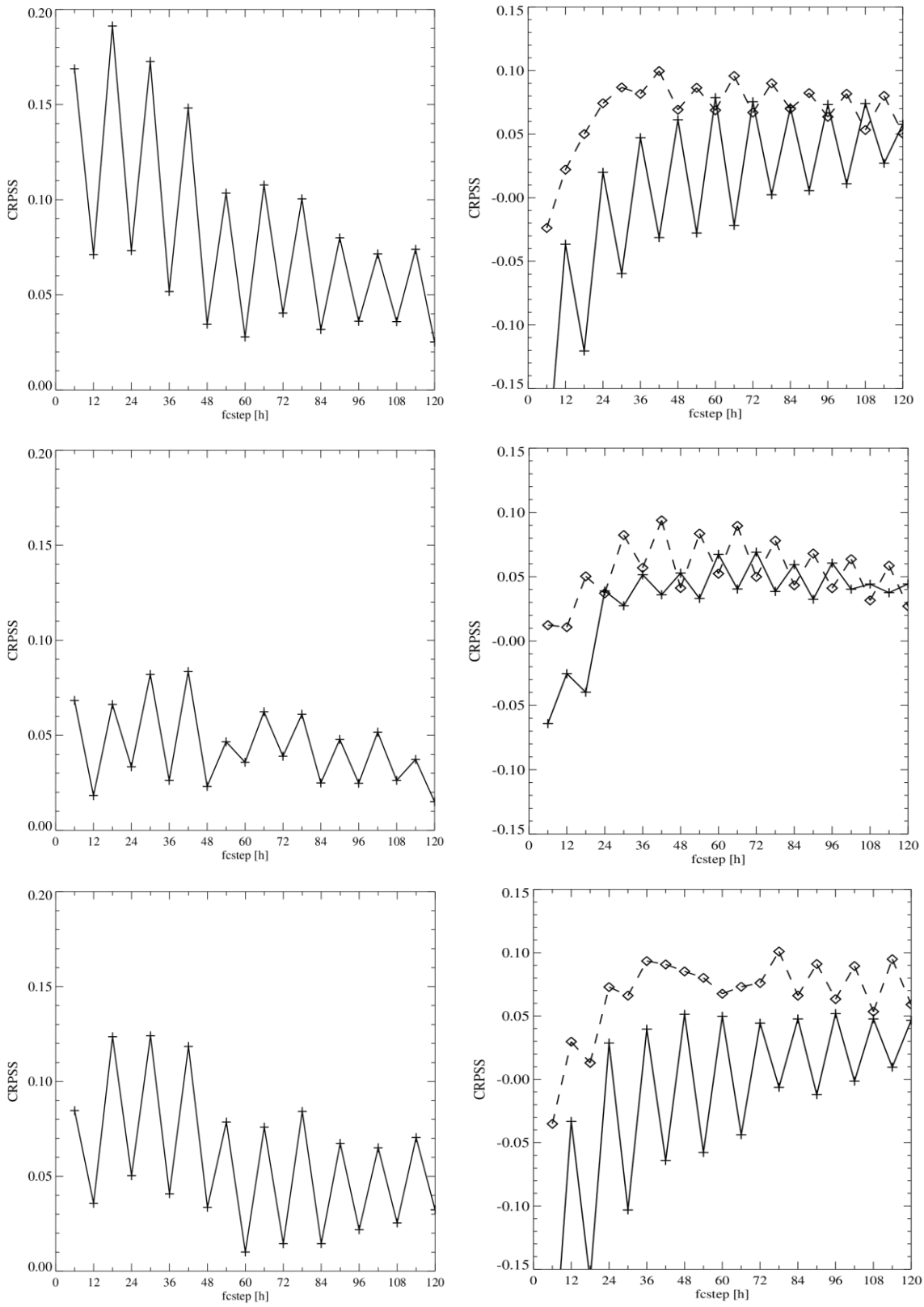


Figure 8.7: *CRPSS for Germany (top), Tennet control zone (middle) and 50Hertz control zone (bottom) utilizing calibrated 10 m Ensemble winds (CPS) to forecast simulated (left) and observed (right) wind power. The reference is the Ensemble Prediction System with uncalibrated winds. A post-processing is applied to remove a strong diurnal bias (dashed line in right figure) in the wind power forecast.*

The 12 h cycle is also existent in the CRPSS when only data from the 00 UTC forecast run are used for the verification (not shown here). Thus, it can be speculated that the 12 h cycle is introduced during the ensemble calibration as forecasts with the same lead time but from different model runs (initialisation at 00 UTC or 12 UTC) have been used. This leads to the effect that the calibration for a given lead time must be valid for two different hours of the day, i.e. the calibration for a given lead time shall commonly account for atmospheric effects at noon and at midnight. Since the atmospheric conditions (mainly thermal stratification of the atmosphere) are substantially different at noon and midnight it is impossible that the calibration works properly. This assumption is fostered by the fact that for the Tennet control zone (Figure 8.7, middle) the double diurnal cycle is weakest compared to the other control zones. It is specific for the Tennet control zone in comparison with the other control zones that the share of far onshore wind power capacity is comparably low and a lot of wind power is located near the coast where diurnal effects of variable wind power are less pronounced.

In general, the improvement in probabilistic skill is higher when verified against simulated wind power compared to observed wind power. Despite strong variations with lead time, the improvement is always positive, but decreasing with increasing lead time. The improvement for verification against simulated wind power is highest for entire Germany.

The verification with observed wind power clearly reveals the deficits of 10 m winds for wind power forecasting since a negative CRPSS up to forecast step +24 h occurs when no wind power bias correction is applied. The negative skill of the CPS for low forecast steps can be explained by a tremendous lack of spread and the inability to forecast low and high wind power properly as can be seen in the Talagrand Diagram in Figure 8.6 (top right). The wind power bias correction improves the ability of CPS to forecast low and high wind power events.
After correcting the diurnal cycle with a post-processing (wind power bias correction), almost at all lead times the CRPSS for the calibrated winds is positive demonstrating an improvement over the raw ensembles. For Tennet the bias correction has the smallest effect and even increases the diurnal variation in CRPSS. CPS improves the probabilistic skill for Germany up to 10 % compared to the uncalibrated 10 m wind power ensemble forecasts.

8.6 Evaluation of extremes

CRPS or CRPSS is not bounded to certain threshold for certain events and consequently not ideal to assess if the calibrated ensemble (CPS) can predict extreme events better. In the following the Brier Score (BS) for events exceeding high (extreme) thresholds is utilized to assess the skill predicting extremes. For comparison with the uncalibrated 10 m ensemble the Brier Skill Score (BSS) is computed as $1\text{-}BS_{CPS}/BS_{ref}$.
As threshold for wind power events 80 % and 70 % installed wind power capacity have been chosen. 80 % and 70 % of installed wind power can already regarded as extreme wind power penetration for Germany as those events occur only for about 22 h and 140 h per year, respectively (Figure 8.8). In the 50Hertz and Tennet control zone the probability of extreme wind power penetration is slightly higher.

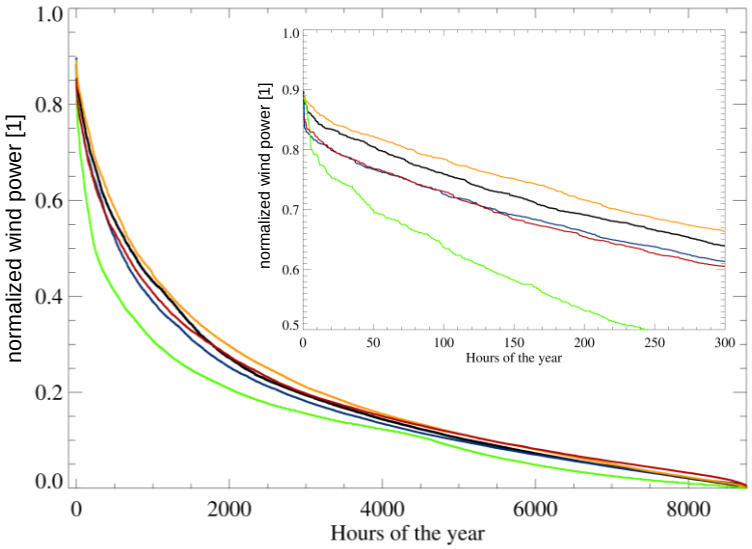


Figure 8.8: *Duration curve of wind power production in control zones of 50Hertz* (black), Amprion (blue), ENBW (green), Tennet (orange) and *Germany (red) for the years 2008 and 2009. The wind power production is smoothed from 15 minute resolution to hourly values.*

The BSS in Figure 8.9 is mostly positive for all events, lead times and areas (control zones) indicating that the CPS leads to improved probabilistic skill when forecasting extreme events. The highest improvement is gained for the Amprion control zone.
In all cases, the daily CRPSS is also positive showing that over the entire wind power spectrum the CPS leads to a more skilful forecast.

The illustration of the increased probabilistic skill of the CPS with a real example of an extreme event is illustrated in Figure 8.10 for the raw ensemble (left) and the calibrated (CPS) ensemble (right). The base time is 22 March 2009 (0 UTC) and the ensemble mean and the deterministic forecast indicate that a major wind power event is expected for Germany. Indeed, the wind power production increased rapidly and exceeded 80 % of installed capacity on 24 March 2009 caused by storm front ‘Herbert’. It can be noted that for forecast steps 18-42 the 50 % inner quantile prediction interval for CPS is slightly shifted and is better centred with respect to observed wind power (red line), i.e. the ensemble mean matches the observed wind power almost perfectly. The 50 % inner quantile prediction interval is reduced at forecast steps 60 and 66 h, i.e. during the period of the decaying storm. The reliable prediction of decreasing wind power after a storm is of major interest when ramping up conventional power plants.

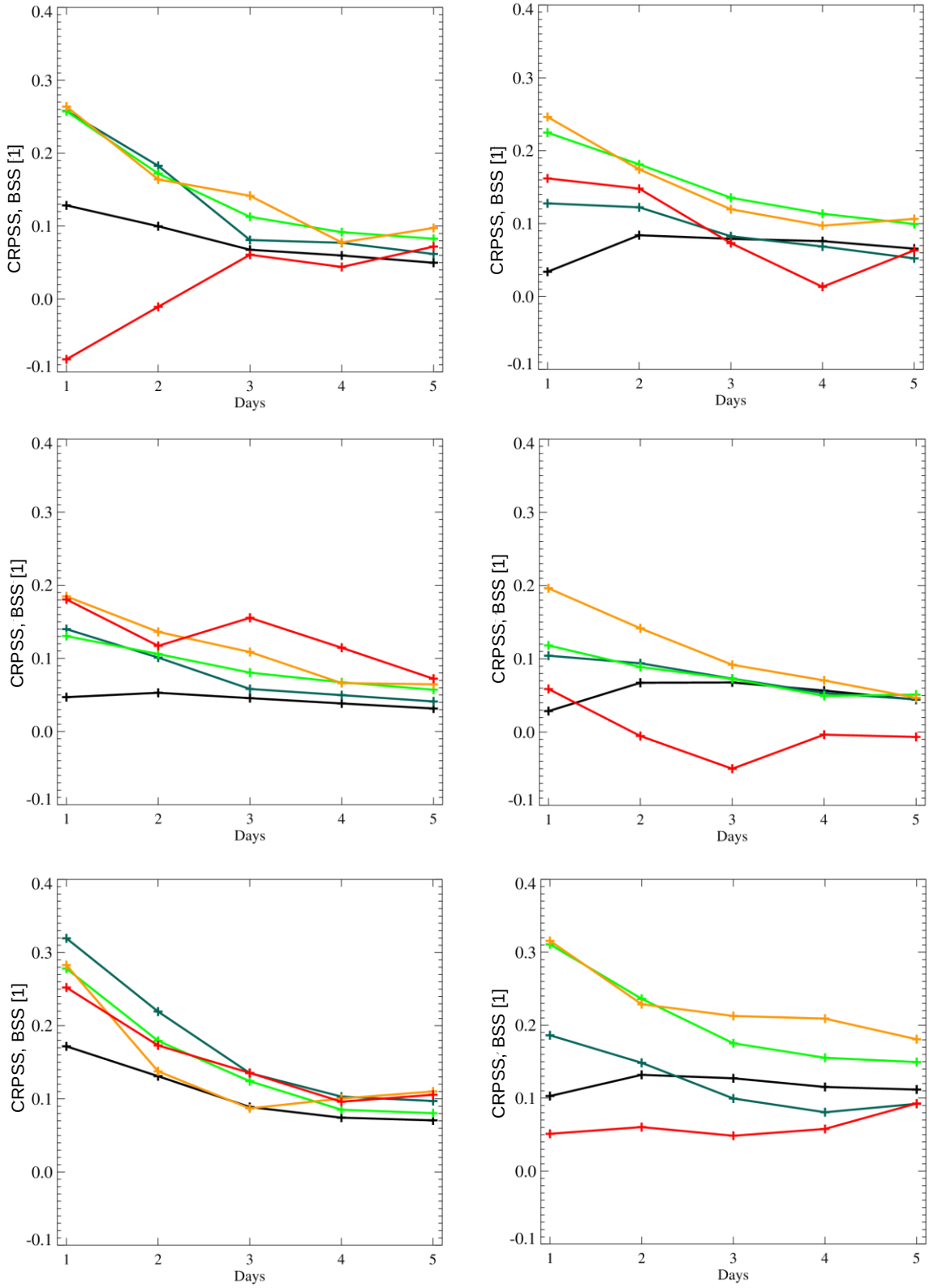


Figure 8.9: *CRPSS (black) for Germany (top), Tennet control zone (middle) and Amprion control zone (bottom) utilizing calibrated 10 m Ensemble winds (CPS) to forecast simulated (left) and observed (right) wind power. The reference is the Ensemble System with uncalibrated winds. A post-processing is applied to remove a strong diurnal bias in the wind power forecast when verified with observed wind power (right). The coloured lines are for the Brier Skill Score (BSS) for events of wind power exceeding 30 % (blue), 50 % (green), 70 % (orange) and 80 % (red) of installed wind power capacity.*

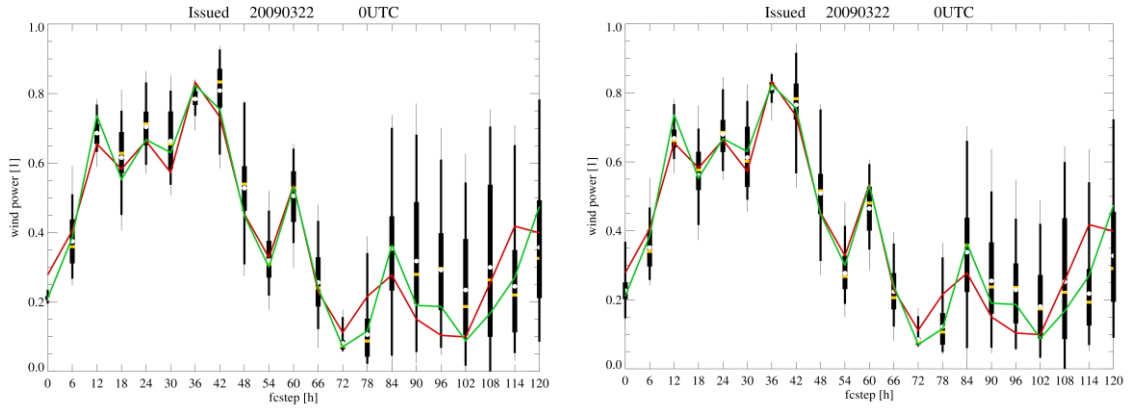


Figure 8.10: *Meteogram of probabilistic German wind power forecast normalized with installed capacity issued at 22 March 2009 (0 UTC) using the raw ensemble (left) and the calibrated 10 m (CPS) ensemble (right). Measured wind power (red), ensemble mean (white dots and deterministic forecast (green). 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.*

9. Results and discussion

The original approach described for the calibration of ensemble forecasts of (u, v)-wind [5] relies on an adaptive and recursive estimation of the parameters of mean and variance models in a ML (Maximum Likelihood) framework. It has the advantage of having solid theoretical foundations while being computationally cheap. This is since model parameters are updated based on the last forecasts and analysis only, every time this analysis is made available. BMA (Bayesian Model Averaging) calibration found to be beneficial concerning a set of ECMWF (raw) ensembles spanning a period of two-year data (2008 & 2009) valid for both the 00 and 12 UTC base time. Results here refer to the full spectrum of wind events.

Results concerning reliability for different wind speed categories for Germany are:

- for 5 m/s: almost perfect reliability for both short- (T+12 to T+60) and early medium-range (T+72 to T+120).

- for 10 m/s: almost perfect reliability until T+96, almost in improvement for T+120

- for 15 m/s: considerable improvement till T+72 hours without reaching perfect reliability. No improvements for higher lead times.

- for 17.5 m/s: superiority CPS till T+48 hours without reaching perfect reliability. No improvements for higher lead times.

Besides reliability diagrams, Talagrand rank (bin) histograms were utilised for the skill assessment of both the raw and calibrated ensembles. Studying the full set of diagrams contained in Appendixes I to W the main results are (concerning the full spectrum of wind speed events):

- Calibration corrects efficiently the U shape (quite pronounced in all raw ensembles especially in the short range) for Germany and Europe. In the early medium-range left outliers are reduced better than right outliers, i.e. sometimes the ensemble members are too low to capture the observed wind speed value.

- No major differences of calibration benefit were found for 00 & 12 UTC forecasts.

Talagrand rank (bin) histograms were also utilised for the skill assessment of both the raw and calibrated ensembles but this time focusing on the extreme events (15 m/s, 17.5 m/s and 20 m/s). Results showed that calibration reduces the population of left outliers for all lead times and extreme event categories very well. Problems are identified for right outliers that are more interesting in the SafeWind context because in their case extreme wind speeds occurred and might have caused damage but have not been forecasted by a single ensemble member. In the short-range CPS is able to decrease the population of right outliers, but beyond T+72 CPS tends to increase the number of right outliers compared to the raw ensemble for all extreme categories. This means that beyond T+72 CPS reduces the ensemble spread for extreme winds.

The general reduction in ensemble spread by CPS can be seen in Figure 8.2 (right) for the evaluation in wind power space for Germany. When verifying CPS wind power forecast against simulated German wind power production a clear advantage over the raw ensemble can be noticed looking at the Talagrand histogram in Figure 8.3. The positive bias in the raw ensemble is efficiently reduced for all lead times and the reliability is clearly improved (Figure 8.4, left).
The verification against real observed wind power in Germany arise the problem that wind speed forecasts in 10 m are not optimal for wind power forecasting. In particular, as no information about the thermal stability of the atmosphere is available a huge wind power bias is introduced by using the logarithmic wind profile for extrapolation to hub height. In a post-processing step this (mainly) time of the day dependent bias is (partly) removed from the wind power forecast and emphasize the superiority of CPS over the raw ensemble in terms of CRPSS improvement (~10 %) and reliability. Nevertheless, this wind power bias correction prevents the CPS ensemble to capture very low wind

power production in the early medium-range (Figure 8.5, bottom right). It must be clearly stated that the wind power bias correction only mimics to account for stability effects as on average the atmosphere is more stable during the night and less stable during the day. However, the following meteorological situation is possible: the wind speed in hub height will be strongly overestimated if a low pressure system (low thermal stability) passes during the night. In SafeWind Deliverable Dp-5.10 [56] it is shown that no wind power bias correction is required when 100 m ensemble winds are used. The superiority of 100 m ensemble winds over 10 m winds is about 25 % in the short-range and 10 % in the early medium-range.

CPS wind power forecasts have a clear better forecast skill for extreme wind power penetrations. This is demonstrated for various German control zones and different penetration levels utilizing the Brier Skill Score (Figure 8.9). The superiority of CPS leading to smaller inner quantiles and better matching the observed wind power is demonstrated for a specific high wind power penetration situation in Germany.

10. Executive summary

The skill of ensemble forecasts as generated by the ECMWF integrated forecast system can be maximised by correcting for their lack of sufficient reliability. A bivariate calibration of these ensemble forecasts has been performed utilising adaptive and recursive estimation of the parameters of mean and variance models in a maximum likelihood framework. The originality of this methodology lies in the fact that calibrated ensembles still consist of a set of (space-time) trajectories, after translation and dilation. An adaptive calibration of ECMWF ensemble forecasts of (u, v)-wind at 10 metres above ground level was applied for Europe over a 3-year period between December 2006 and December 2009.

The results have shown that 10 m calibrated ensemble winds of the Calibrated Prediction System (CPS) developed at ECMWF within SafeWind are beneficial for both the short- and early medium-range concerning a set of ECMWF (raw) ensembles spanning a period of two-year data (2008 & 2009). Results in wind speed mode refer to the assessment of reliability for non-extreme wind events verified against ECMWF wind analyses. In such cases, almost perfect reliability results and flat Talagrand histograms can be obtained by CPS compared to the raw ensemble.

For extreme wind events, the superiority of the CPS over the raw ensemble is limited to the short-range when verified against wind analyses. Beyond the short-range no improvements with respect to reliability and capturing of extremes (as assessed by the highest rank in the Talagrand histogram) are obtained.

German wind power forecasts can be substantially improved with the CPS ensemble in terms of the Continuous Rank Probability Skill Score and in terms of the Brier Skill Score for the forecast of extreme wind power events. However, deficiencies in the spread of CPS (and raw) ensembles have been observed that are related to a strong wind power biases introduced by the required extrapolation of 10 m winds to hub height of wind turbines Thus, the forecasting of extremes in wind power (high and low wind power production) is unfortunately slightly impaired.

It is important to note that calibration cannot solve all problems relating to an improvement of skill, especially when focusing on extremes. Such limitations have been observed in similar cases [53] when improvements by Bayesian Model Averaging (BMA) calibration were smaller for extreme events.

From a more general point of view, a comment regarding the role of a meteorological forecast provider who calibrates ensemble forecasts with respect to its own target (analysis) is made. It is obvious that for forecast users the actual target may be different and depending upon the intended application e.g. in the case of local observations (measurements) of wind speed and direction at a wind farm. Hence even if ensembles are calibrated perfectly with respect to their own target (analysis) by the forecast provider, further calibration is likely to be necessary before ensemble/probabilistic forecasts are to be used in decision-making.

11. References

[1] Pinson P., 2012: Adaptive calibration of (u, v)-wind ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 138 (666): 1273-1284.

[2] Bao L., Gneiting T., Grimit E.P., Guttorp P. and A.E. Raftery, 2010: Bias correction and Bayesian model averaging for ensemble Forecasts of surface wind direction. Monthly Weather Review 138: 1811–1821.

[3] Sloughter J.M., 2009: Probabilistic weather forecasting using Bayesian model averaging. PhD final presentation, University of Washington, Dpt. of Statistics, Seattle (USA).

[4] Sloughter J.M., Gneiting T. and A.E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. Journal of the American Statistical Association 105: 25–35.

[5] Pinson P., 2011: Calibrated ensemble forecasts of (u,v)-wind. SafeWind's combined Deliverable 5.6 & 5.7 (final report).

[6] Casati B., Wilson L.J., Stephenson D.B., Nurmi P., Ghelli A., Pocernich M., Damrath U., Ebert E.E., Brown B.G. and S. Mason, 2008: Review, forecast verification: current status and future directions. Meteorol. Appl. 15, 3-18.

[7] Murphy A.H., 1991: Probabilities, odds and forecasts of rare events. Wea. Forecasting 6, 302-307.

[8] WMO brochure on extreme events 2001-2010, 2011: Weather extremes in a changing climate – Hindsight on foresight. *http:// www.wmo.int/pages/mediacentre/news/documents/1075_en.pdf.*

[9] Bougeault P., 2003: The WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. Meteo France, Toulouse, France.

[10] Legg T.P. and K.R. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. Weather Forecasting, 19, 891-906.

[11] Gall R. and M. Shapiro, 2000: The influence of Carl-Gustaf Rossby on mesoscale weather prediction and an outlook for the future. Bull. Amer. Meteor. Soc., 81, 1507-1523.

[12] Rabier F., Thépaut J.-N. and P. Courtier, 1998: Extended assimilation and forecast experiments with a four- dimensional variational assimilation system. Q.J.R. Meteor. Soc., 124, 1861-1887.

[13] Hello G., Lalaurette F. and J.-N. Thépaut, 2000. Combined use of sensitivity information and observations to improve meteorological forecasts: A feasibility study applied to the 'Christmas Storm' case', Q.J.R. Meteor. Soc., 126, 621-647.

[14] Joly A., Browning K.A., Bessemoulin P., Cammas J.-P., Caniaux G., Chalon J.-P., Clough S. A., Dirks R., Emanuel K. A., Eymard L., Gall R., Hewson T. D., Hildebrand P. H., Jorgensen D., Lalaurette F., Langland R. H., Lemaitre Y., Mascart P., Moore J. A., Persson P. O., Roux F., Shapiro M. A., Snyder C., Toth Z. and Wakimoto R. M., 1999. Overview of the field phase of the Fronts & Atlantic Storm-Track EXperiment (FASTEX) project. Q.J.R. Meteor. Soc. 125 877-946.

[15] Lalaurette F. and G.v.d. Grijn, 2005: Predictability of Weather and Climate: Ensemble forecasts: can they provide useful early warnings? In Palmer T.N. and R. Hagedorn, editors, Predictability of Weather and Climate. Cambridge University Press.

[16] Gneiting T. and A.E. Raftery, 2005: Weather forecasting with ensemble methods. Science 310: 248–249.

[17] Palmer T.N., 2000: Predicting uncertainty in forecasts of weather and climate. Reports on Progress in Physics 63: 71–116.

[18] Nielsen HAa., Nielsen T.S., Madsen H., Giebel G., Badger J., Landberg L., Sattler K., Voulund L. and J. Tøfting, 2006: From wind ensembles to probabilistic information about future wind power production - results from an actual application. In Proceedings of 2006 IEEE PMAPS Conference, Probabilistic Methods Applied to Power Systems, Stockholm, Sweden.

[19] Pinson P. and H. Madsen, 2009: Ensemble-based probabilistic forecasting of wind power at Horns Rev. Wind Energy 12: 137–155.

[20] Taylor J.W., McSharry P. and R. Buizza, 2009: Wind power density forecasting using ensemble predictions and time- series models. IEEE Transactions on energy conversion 24: 775–782.

[21] Matos M.A. and R. Bessa, 2010: Setting the operating reserve using probabilistic wind power forecasts. IEEE Transactions on Power Systems, in press.

[22] Meibom P., Barth R., Hasche B., Brand H., Weber C. and M. O'Malley, 2010: Stochastic optimization model to study the operational impacts of high wind penetrations in Ireland. IEEE Transactions on Power Systems, in press.

[23] Pinson P., Chevallier C. and G. Kariniotakis, 2007: Trading wind generation from short-term probabilistic forecasts of wind power. IEEE Transactions on Power Systems 22: 1148–1156.

[24] Gneiting T., 2011: Quantiles as optimal point predictors. International Journal of Forecasting 27: 197–207.

[25] Thorarinsdottir T.L. and T. Gneiting, 2010: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society, Series A 173: 371–388.

[26] Gneiting T., Stanberry L., Grimit E.P., Held L. and N.A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17: 211–235.

[27] Wilks D.S., 2002: Smoothing forecast ensembles with fitted probability distributions. Quarterly Journal of the Royal Meteorological Society 128: 2821–2836.

[28] Meibom P., Barth R., Hasche B., Brand H., Weber C. and M. O'Malley, 2010: Stochastic optimization model to study the operational impacts of high wind penetrations in Ireland. IEEE Transactions on Power Systems, in press.

[29] Morales J.M., Conejo A.J. and J. Pérez, 2010: Short-term trading for a wind power producer. IEEE Transactions on Power Systems 25 554–564.

[30] Buizza R., Richardson D.S. and T.N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. Quarterly Journal of the Royal Meteorological Society 129: 1269–1288.

[31] Vannitsem S. and R. Hagedorn, 2010: Ensemble forecast post-processing over Belgium: comparison of deterministic- like and ensemble regression approaches. Meteorological Applications, available online.

[32] Gneiting T., Raftery A.E., Westwald A.H. III and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133: 1098-1118.

[33] Simmons A., Burridge D.M., Jarraud M. and W. Wergen, 1988: The ECMWF medium-range prediction models; Development of the numerical formulations and the impact of increased resolution. Meteorol. Atmos. Phys., 40: 28-60.

[34] Palmer T.N., Molteni F., Mureau R., Buizza R., Chapelet P. and J. Tribbia, 1993: Ensemble Prediction. In proceedings 1992 ECMWF Seminar: Validation of Models over Europe, pp 21-66. ECMWF, Reading, U.K.

[35] Molteni F., Buizza R., Palmer T. N. and T.I. Petroliagis, 1996: The new ECMWF ensemble prediction system: methodology and validation. Q.J.R. Meteor Soc., 122, 73-119.

[36] Palmer T.N., Buizza R., Leutbecher M., Hagedorn R., Jung T., Rodwell M., Vitart F., Berner J., Hagel E., Lawrence A., Pappenberger F., Park Y-Y., Bremen v.L. and I. Gilmour, 2007: The Ensemble Prediction System – Recent and Ongoing Developments. Tech. Memo 540. ECMWF, Reading, U.K.

[37] Buizza R., Houtekamer P.L., Toth Z., Pellerin G., Wei M. and Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. Monthly Weather Review 133: 1076–97.

[38] Leutbecher M. and T.N. Palmer T.N., 2008: Ensemble forecasting. Journal of Computational Physics 227: 3515–3539. Madsen H. Time Series Analysis, Chapman & Hall/CRC: London, 2007.

[39] Magnusson L., Leutbecher M. and E. Kallén, 2008: Comparison between singular vectors and breeding vectors as initial perturbations for the ECMWF ensemble prediction system. Monthly Weather Review 136: 4092–4104.

[40] Buizza R., Miller M. and T.N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 131: 2887–2908.

[41] Palmer T.N., Shutts G.J., Hagedorn R., Doblas-Reyes F.J., Jung T. and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. Annual Review of Earth and Planetary Sciences 33: 163–193.

[42] Hering A.S. and M.G. Genton, 2009: Powering up with space-time wind forecasting. Journal of the American Statistical Association 105: 96–104.

[43] Stone M., 1974: Cross-validation and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, Series B 36: 111–147.

[44] Leung S.-H. and C.F. So CF, 2005: Gradient-based variable forgetting factor RLS algorithm in time-varying environment. IEEE Transactions on Signal Processing 53: 3141–3150.

[45] Paleologu C., Benesty J. and S. Ciochina, 2008: A robust variable forgetting factor Recursive Least-Squares algorithm for system identification. IEEE Signal Processing Letters 15: 597–600.

[46] Hartmann H.C., Pagano T.C., Sorooshiam S. and R. Bales, 2002: Confidence builder: evaluating seasonal climate forecasts from user perspectives. Bull Amer. Met. Soc., 84, 683-69.

[47] Lalaurette F., 2002: Early Detection of Abnormal Weather Using a Probabilistic Extreme Forecast Index. Q.J.R. Meteor Soc. (also available as ECMWF Technical Memo No. 373).

[48] Petroliagis T.I. and P. Pinson, 2011: Early warnings and alerts of extreme wind events utilising DWD objective weather type classification methodology and ECMWF EPS Extreme Forecast Index. SafeWind's Deliverable 5-5 (final report). Also submitted to RMetS Meteorological Applications.

[49] Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. Mon. Wea. Rev., 129, 550–560.

[50] Hamill, T.M. and S.J. Colucci, 1997: Verification of Eta–RSM short- range ensemble forecasts. Mon. Wea. Rev., 125, 1312–1327.

[51] Hamill, T.M. and S.J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. Mon. Wea. Rev., 126, 711–724.

[52] Eckel, F. A, and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. Wea. Forecasting, 13, 1132-1147.

[53] Iversen T., Deckmyn A., Santos C., Sattler K., Bremnes J.-B., Feddersen H. and I.-L. Frogner, 2001: Evaluation of “GLAMEPS” - a proposed multi-model EPS for short-range forecasting, Tellus A, 63A, 513-530.

[54] Betreiber-Datenbasis, www.btrdb.de

[55] McLean JR, 2008: Equivalent Wind Power Curves. Deliverable 2.4 of the TradeWind Project

[56] Deliverable 5.10 of the SafeWind Project, August 2012

[57] Broecker, J., L. A. Smith, 2007: Increasing the reliability of reliability diagrams. Weather and Forecasting, 22: 651-661

[58] Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. International Geophysics Series, 3rd edition, 2011.