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 Dc-5.11b
"Progress during the second half of the project and future prospects"

DOCUMENT TYPE	Deliverable
DOCUMENT NAME:	swind.deliverable_Dc-5.11b_v2_0.doc
VERSION:	v2.0
DATE:	2012.10.19
CLASSIFICATION:	R0: General Public
STATUS:	Approved / Released

Abstract:

The verification of CPS wind power forecasts against real observed wind power in Germany arise the problem that wind speed forecasts at 10 meters are not optimal for wind power forecasting. Furthermore, the required wind power bias correction prevents the CPS ensemble to capture very low wind power production in the early medium-range. CPS wind power forecasts have better forecast skill for extreme wind power penetrations. Furthermore, it has been demonstrated that the new EPS product of ECMWF to archive and disseminate winds at 100-meter height is a major achievement for probabilistic wind power forecasting. No wind power bias correction is necessary since the 100-meter winds can be directly interpolated to the hub height of modern wind turbines. This leads to an improvement of almost 50% in wind power RMSE (for Germany) using the 100 meter ensemble mean compared to the 10-meter ensemble mean. In cases of extremes, the exploitation of the ECMWF Extreme Forecast Index (EFI) based on its 10-meter Gust Factor formulation (as suggested in Dp-5.5) was further investigated. Overall, it became clear that the first indication of an extreme wind event could come from the ability of IFS and EPS components to capture very intense cyclonic circulation systems (reflecting to possible windstorms) in the medium- and even late medium-range. Furthermore, from all applications of the EFI schemes, the highest skill in issuing early warnings is given by the EFI-10FGI formulation based on the 00 UTC forecast cycle. Although ROCA (Area under the ROC Curve) values are found to be very high, suggesting a skilful performance, in a real operational mode use of the 99% EFI percentile threshold would provide early warning for a considerable number of >99% category extremes, but not for all. By lowering this threshold the number of hits is increased till all extremes are captured (reflecting to zero misses), but by doing so, the number of false alarms is increased significantly. Consequently, an optimal trade-off between hits and false alarms has to be made when setting various (critical) EFI thresholds. Already the EFI is a key resource for helping forecasters provide warnings of severe weather events. To provide additional assistance, ECMWF has immediate plans to extend the EFI out to 7 days. A set of different examples of how to apply EFI in cases of high-impact events and how to combine its information with other EPS products is documented.

AUTHORS[1], REVIEWERS				
MAIN AUTHOR/EDITOR:	Thomas I. Petroliagis			
AFFILIATION:	ECMWF (European Centre for Medium-Range Weather Forecasts)			
ADDRESS:	Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom			
TEL.:	+44 (0) 118 94 99 661			
EMAIL:	thomas.petroliagis@ecmwf.int			
FURTHER AUTHORS:	Lueder von Bremen, ForWind – University of Oldenburg			
PEER REVIEWERS:	Robin Girard (Armines, France)			
REVIEW APPROVAL:	Approved:	X	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):
v0.0	2012-08-12	First draft focusing on UNIOL contribution	Lueder von Bremen
v0.1	2012-08-27	Extended draft including all latest applications based on the EFI	Thomas I. Petroliagis
v1.0	2012-09-25	Integration & harmonization among contributions from partners	Thomas I. Petroliagis
v2.0	2012-10-19	Final version after revision	Thomas I. 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	**X**
S2	Pending for review			**R2**	Restricted to European Commission			Paper copy	**X**
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

1. Calibrated Prediction System(s) for ensemble wind power prediction

In this work the benefits and limits of the Calibrated Prediction System (CPS) for 10-meter ensemble winds are analysed respect to real observations in wind power for the example of German wind power predictions. The wind power forecasts are computed individually for each of the four control zones and finally aggregated. The study period is 2008 and 2009.

1.1 Wind power forecast model

The used wind power forecasting model for Germany is intentionally kept very basic. The data base [1] 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 database 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 containing 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 7 x 4.2 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 cannot 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 [2]. The power curve is shown in Figure 1. More details about the wind power-forecasting model can be found in [3].

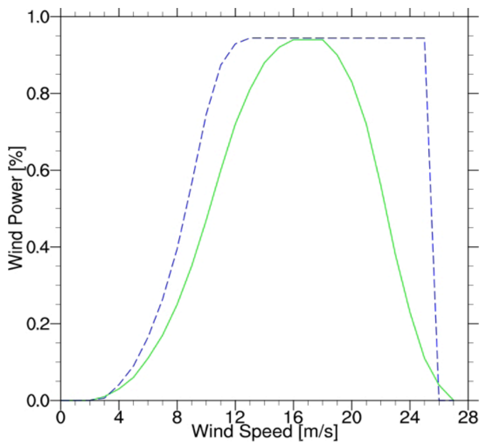


Figure 1: Regional power curve from the TradeWind Project (green) and manufacturer's onshore power curve (blue) for a multi-megawatt wind turbine.

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 i.e. bias (Figure 2, left).

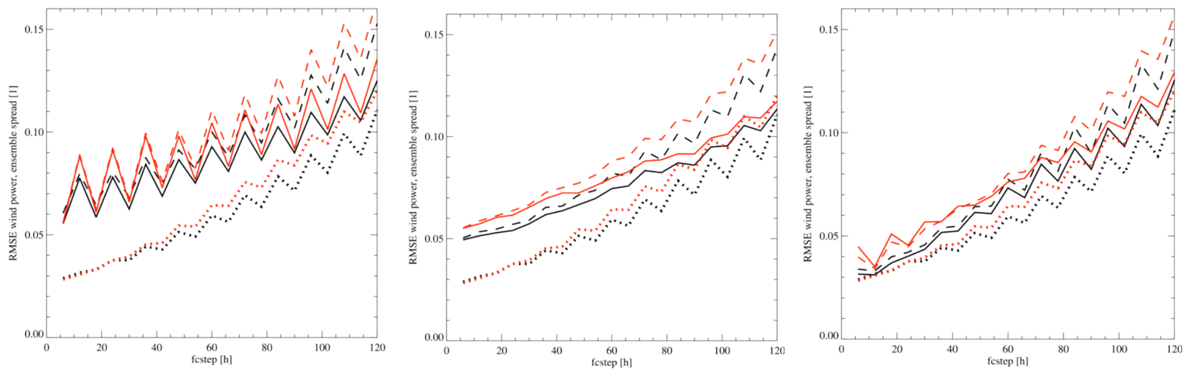


Figure 2: Ensemble wind power forecast spread (dotted line), skill (RMSE) of ensemble mean wind power forecast (full line) and control forecast (dashed line) over forecast step. Wind power forecasts are computed for Germany with raw 10-meter winds (red) and calibrated 10-meter winds (black). Left: verification 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-meter wind speeds are unbiased with respect to the ECMWF analysis.

1.2 Deterministic forecast verification

The (deterministic) forecast skill, expressed as RMSE normalized with the installed capacity, is shown in Figure 2. The wind power bias correction is very efficient to remove the diurnal bias (Figure 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 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.

1.3 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 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 3 for forecast Day+1, +3 and +5. The Talagrand Diagram for the raw ensemble (Figure 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 3, right).

At Day+1 the calibrated ensemble is slightly over-dispersive. 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. 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 4, left). Calibrated forecast probabilities for the event wind power >50% of installed capacity matches 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 [4] 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.

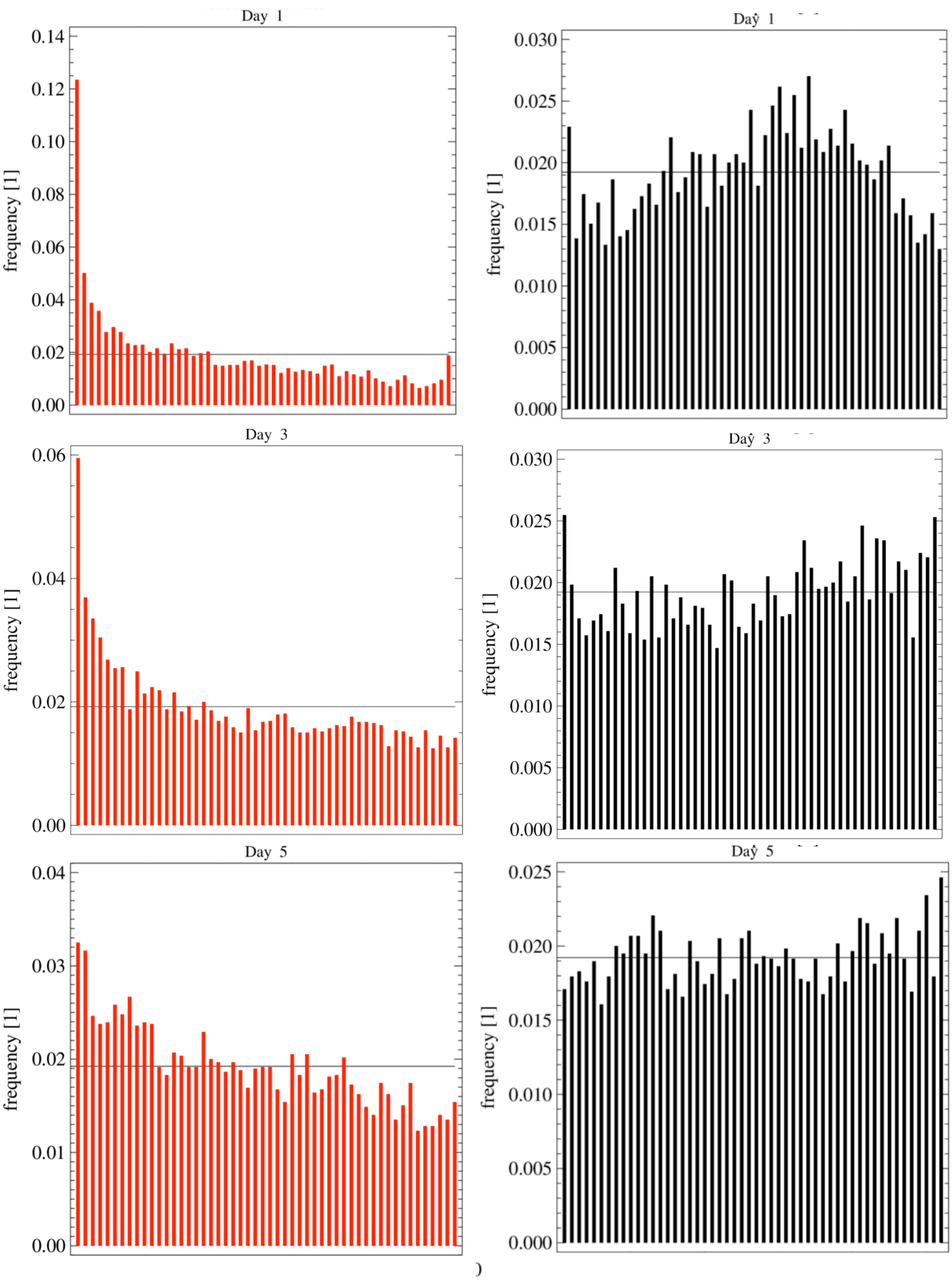


Figure 3: Talagrand Diagram for forecast Day+1, +3 and +5 for uncalibrated (red) and (CPS) calibrated (black) 10-meter EPS winds forecasting German wind power. The verification is done against simulated wind power.

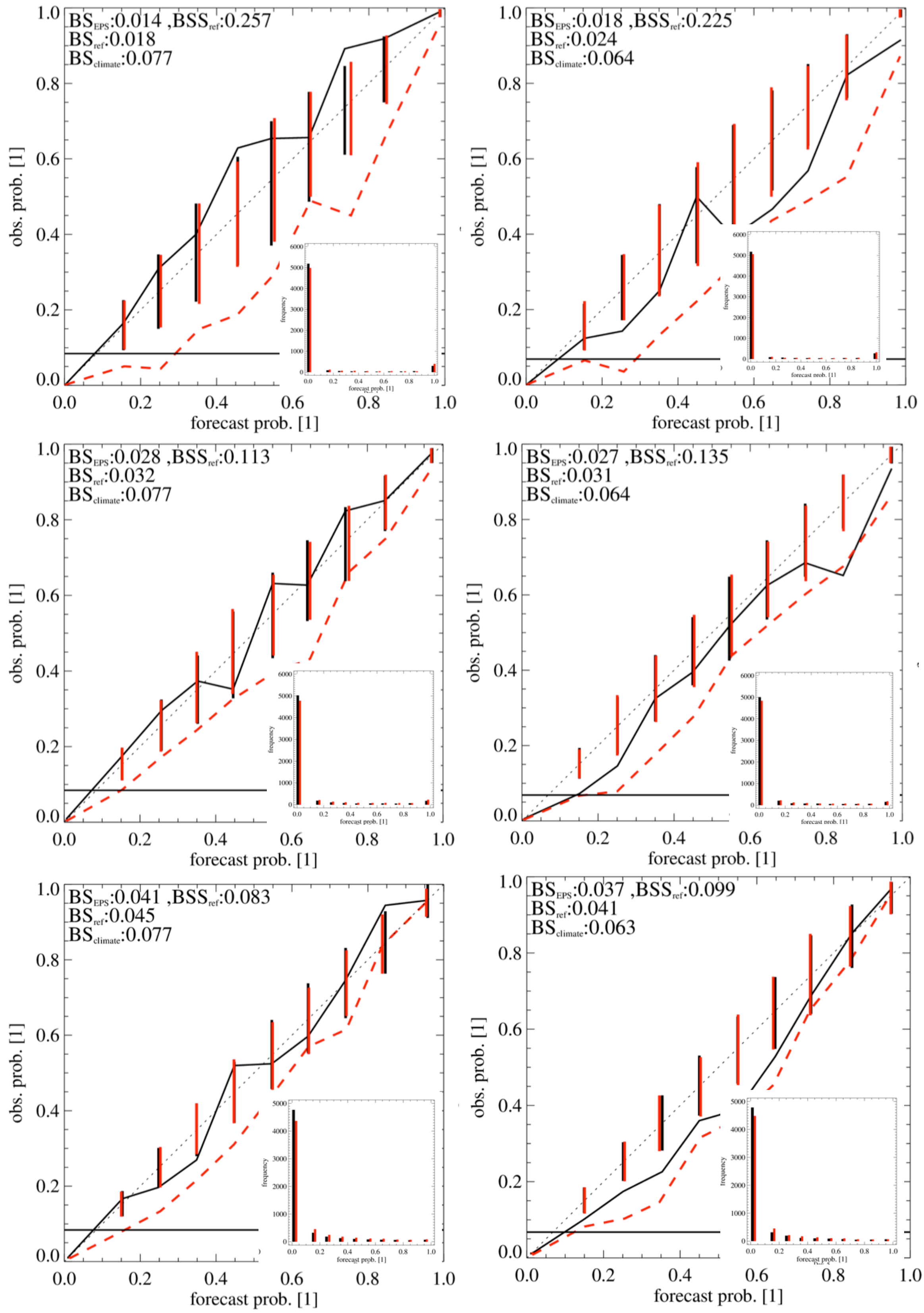


Figure 4: Reliability Diagram for the event wind power >50 % of installed capacity for (CPS) calibrated (black) and uncalibrated (red) 10-meter EPS winds forecasting German wind power at forecast Day+1 (top), Day+3 (middle) and Day+5 (bottom). The verification is done against simulated wind power (left) 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.

1.4 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 5, top) are U-shaped and in contrast to the verification against simulated wind power, no advantage of the CPS winds can be noted.

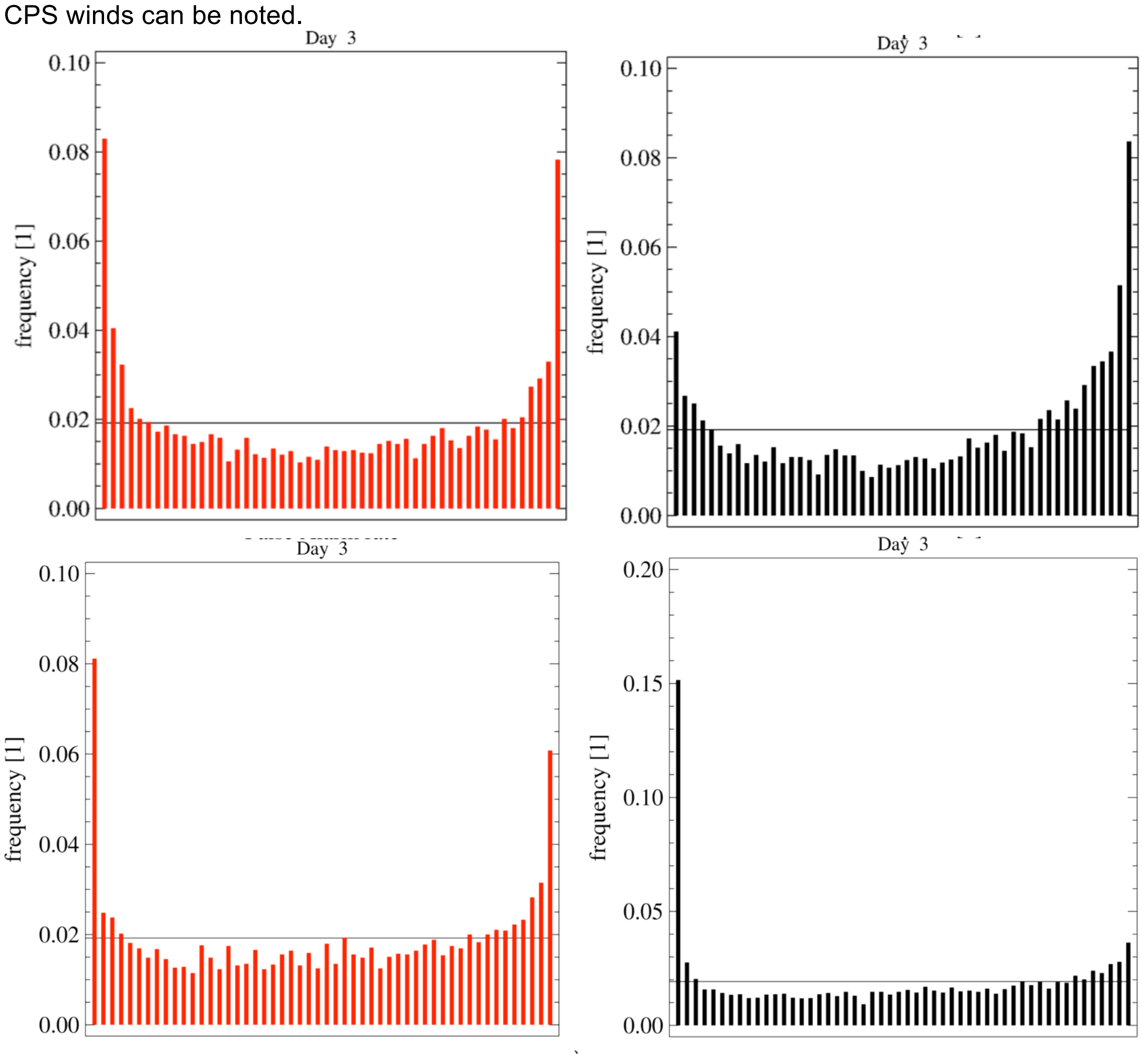


Figure 5: Talagrand Diagram at forecast Day+3 for uncalibrated (red) and calibrated (black) 10 m EPS winds forecasting German wind power. The verification of not biased-corrected (top) and bias-corrected (bottom) wind power forecast is done against observed wind power.

The post-processing to remove the large wind power bias does not change the Talagrand Diagram of the raw ensemble (Figure 5, left), but the Talagrand Diagram for CPS degrades even further (Figure 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 2 (middle) compared to Figure 2 (right) where the verification is done against simulated wind power. The reliability of the calibrated ensemble predicting wind power >50% of installed wind power is clearly improved compared to the raw ensemble (Figure 4, right). The Brier Skill Score (BSS) printed in Figure 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 5 (lower right) is still clearly below the diagonal and outside the consistency bars.

1.5 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 [5].

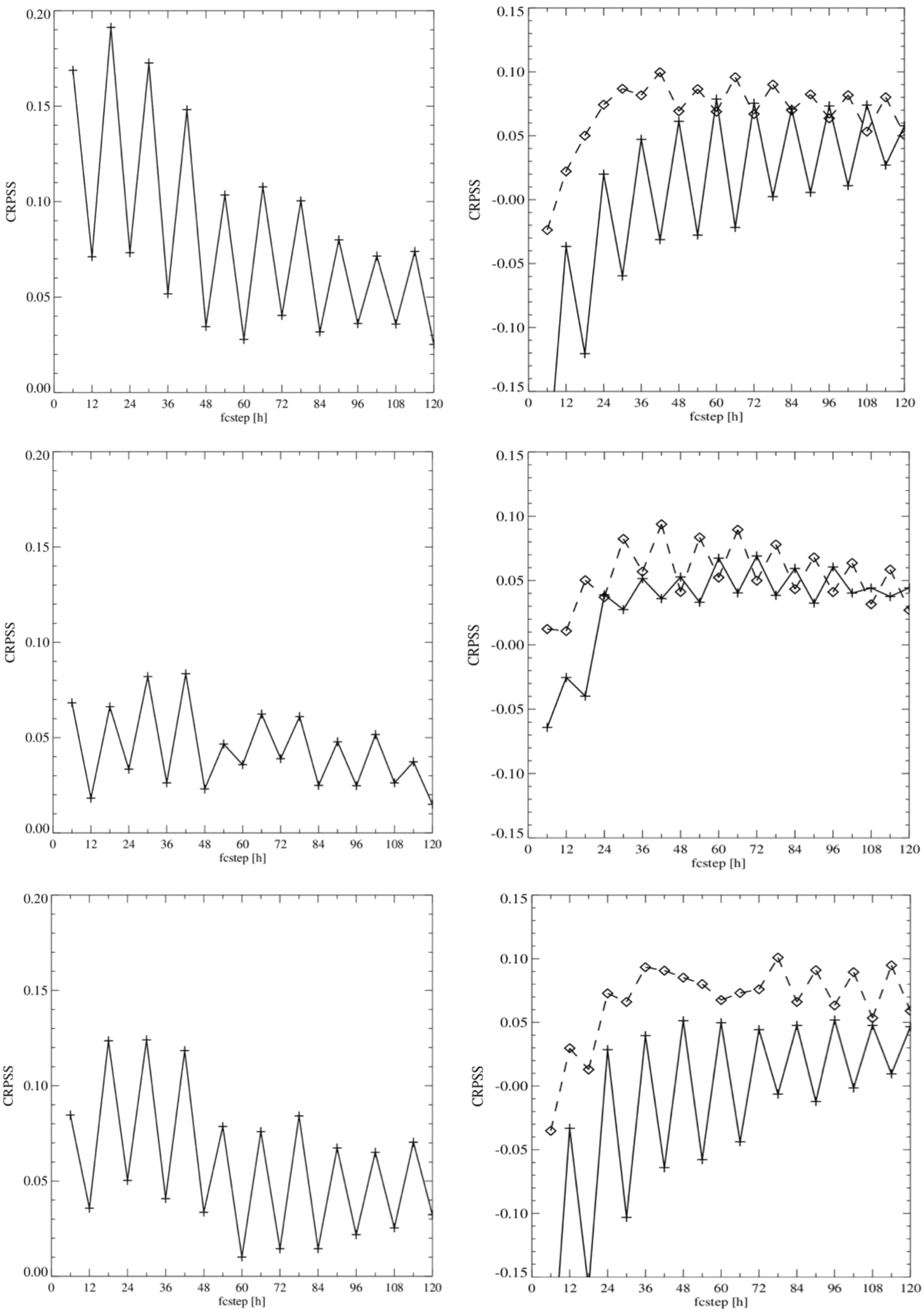


Figure 6: CRPSS for Germany (top), Tennet control zone (middle) and 50Hertz control zone (bottom) utilizing calibrated 10-meter Ensemble winds (CPS) to forecast simulated (left) and observed (right) wind power. The reference is the Ensemble Prediction System with uncalibrated winds. A post-processing is applied to remove a strong diurnal bias (dashed line in right figure) in the wind power forecast.

In the reference ensemble system uncalibrated 10 m winds are used. Figure 6 shows the CRPSS for two control zones (Tennet and 50Hertz) and entire Germany. The verification is done with simulated wind power (from 10-meter 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 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].

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 6, 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 wind power bias correction improves the ability of CPS to forecast low and high wind power events and at almost 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.

1.6 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 (see [6], Figure 8).

The BSS in Figure 7 is mostly positive for all events and lead times indicating that the CPS leads to improved probabilistic skill when forecasting extreme events. 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 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.

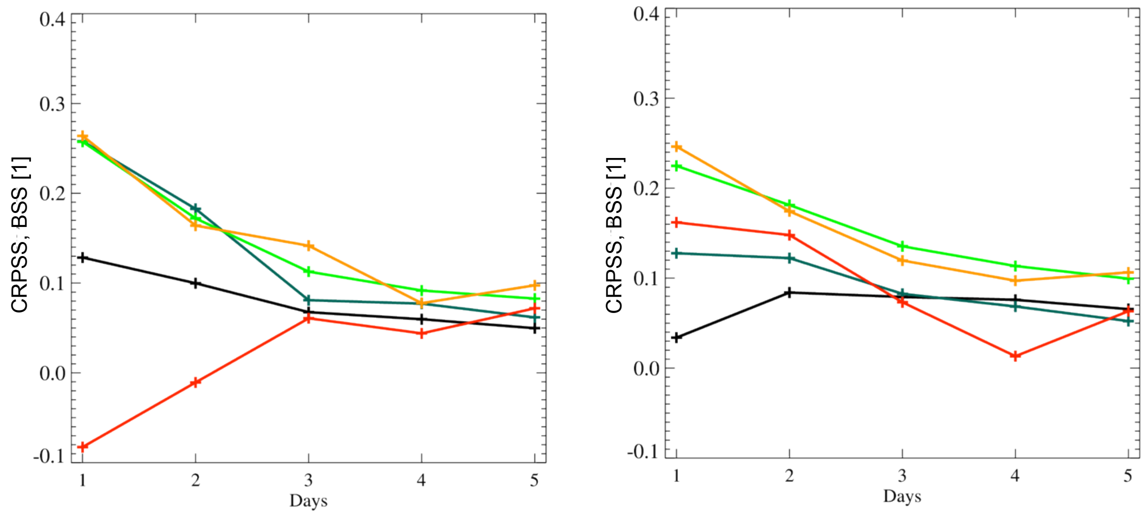


Figure 7: CRPSS (black) for Germany utilizing calibrated 10-meter Ensemble winds (CPS) to forecast simulated (left) and observed (right) wind power. The reference is the Ensemble System with uncalibrated winds. A post-processing is applied to remove a strong diurnal bias in the wind power forecast when 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.

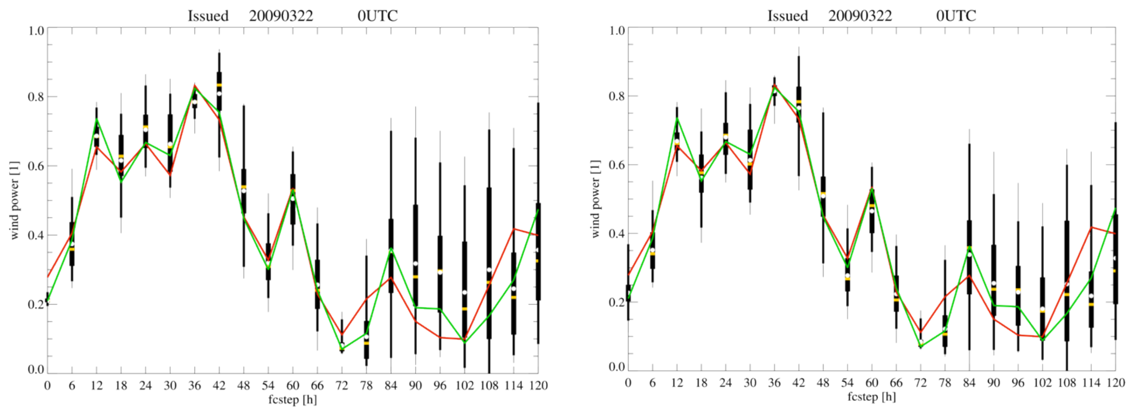


Figure 8: Meteogram of probabilistic German wind power forecast normalized with installed capacity issued at 22 March 2009 (0 UTC) using the raw ensemble (left) and the calibrated 10-meter (CPS) ensemble (right). Measured wind power (red), ensemble mean (white dots and deterministic forecast (green). 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.

2. Enhanced wind power forecasts with 100-meter ensemble winds

2.1 Introduction

Since years the uncertainty in Numerical Weather Prediction (NWP) is quantified by Ensemble Forecasting [7]. At the European Centre for Medium-Range Weather Forecasts (ECMWF) 50 members of the Ensemble Prediction System (EPS) are computed with slightly different initial conditions. The different trajectories of the forecasts can be used to estimate the probability of certain events or to quantify in general if the uncertainty of the forecast is high or low.

Since years wind power forecasters have been correcting the effect of atmospheric stability on winds in hub height. Without stability correction hub height winds are underestimated (overestimated) in stable (non-stable) atmospheric conditions when logarithmically extrapolated from surface winds (10 m height). However, deterministic forecasts provide the required data for the stability correction. This is not the case for ensemble prediction systems due to data volume constraints. Within the SafeWind project ECMWF introduced 100 m winds as a new product tailored for the wind energy industry in the analysis, deterministic and ensemble-forecasting suite at 26 January 2010.

2.2 Benefits from 100-meter ensemble winds

Wind power forecasts are evaluated for Germany interpolating 10 m and 100-meter winds to hub height for each grid point, respectively. A regional power curve from TradeWind is used for the conversion from wind to power [8]. Wind power forecasts with 100-meter winds are clearly superior to 10-meter winds [9]:

- decreased ensemble mean forecast error for Germany up to 50 %
- RMSE forecast error <8 % for forecast Day+3 using ensemble mean
- strong improvement in ensemble spread and reliability
- up to 25 % improvement in probabilistic skill score CRPSS

2.3 Evaluation of probabilistic wind power forecast for Germany

Wind power forecasts have been computed for each of the four control zones in Germany individually. Unfortunately, the study period is limited by the availability of the new 100-meter EPS winds. Thus, the study period is from 26 Jan 2010 until 30 April 2011. Forecasts utilizing 10-meter winds require a bias correction that is dependent on the time of the day as very large biases occur due to dismissed thermal stability effects when extrapolating 10-meter wind speeds to hub height (Figure 9, green and red curve). The Talagrand Rank Histogram for 10-meter winds is not improved by the simple bias correction (Figure 10, left) and shows that low and high wind power events are very often outside the ensemble range when 10-meter wind speeds have been used.

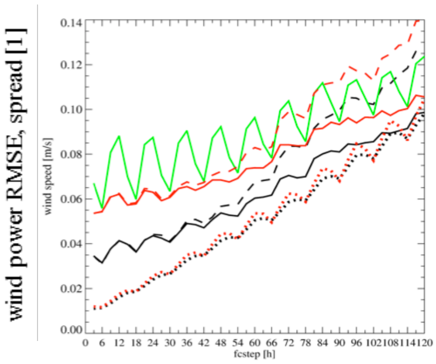


Figure 9: RMSE of wind power forecast normalized with installed capacity for Germany utilizing 100-meter (black) and 10-meter bias corrected (red) ensemble winds (black). The ensemble mean (solid line) clearly outperforms the deterministic forecast (dashed) at forecast Day+3. 10-meter ensemble mean forecast without bias correction (green) and ensemble spread (dotted) is also shown.

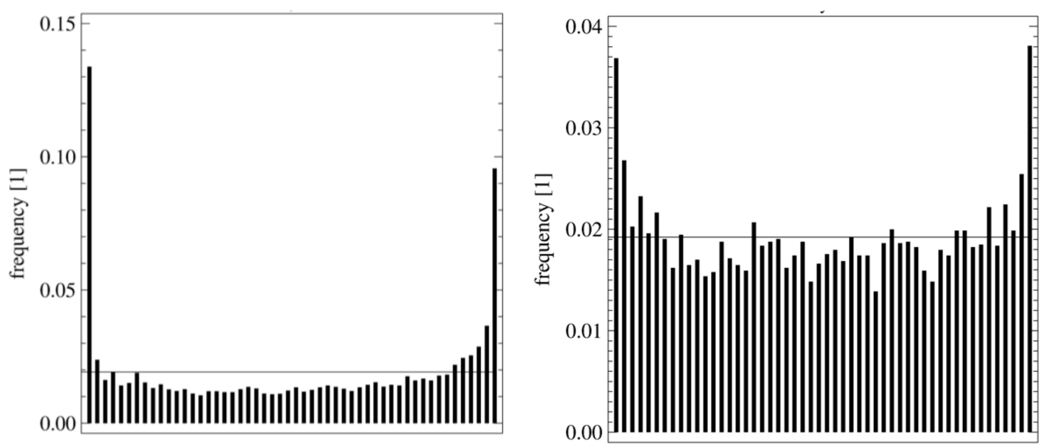


Figure 10: Talagrand Rank Histogram for German wind power at forecast Day+3. Bias corrected 10 m ensemble winds (left) and 100 m ensemble winds (right).

Bias correction is not needed for 100-meter wind power forecasts. Compared to 10-meter winds the Talagrand Rank Histogram looks very much improved, i.e. the spread of 100-meter wind power forecast members is considerably better. The criteria for a skilful probabilistic forecast (reliability, sharpness, resolution) are combined in the Continuous Ranked Probability Score. The comparison with a reference probabilistic forecast system leads to the Continuous Ranked Probability Skill Score (CRPSS). The reference system is outperformed when CRPSS is larger than zero.

The superiority of 100 m over 10m ensemble winds decreases with increasing forecast step (Figure 11). The improvement for the control zone of 50Hertz is slightly smaller than for Germany.

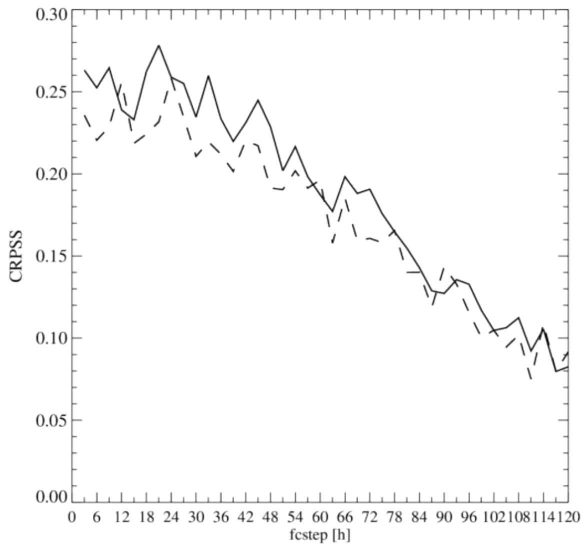


Figure 11: Improvement in skill (CRPSS) for an ensemble wind power prediction system for Germany (full line) and 50Hertz (dashed line) utilizing 100-meter ensemble winds. The reference ensemble system uses 10-meter ensemble winds. The time period is Feb 2010 to Apr 2011.

2.4 Evaluation of probabilistic wind power forecast for a French wind farm

The impact of ensemble model level winds on wind power forecasts for a French wind farm was analysed with respect to the usage of 10-meter ensemble winds. The wind power forecasts were computed with a Neural Network. It is found that forecasts obtained from model levels closest to hub height (78 meters) and the linear interpolation between adjacent levels have the best probabilistic skill (Figure 12).

In general, the improvements in CRPSS are less compared to Germany. It will be shown later if 100-meter winds have the highest probabilistic relevance for large regions (e.g. Germany).

The calibration of 10-meter winds [10] does not improve the CRPSS substantially for a single site. However, it is expected that the calibration of raw model level winds or 100-meter winds will improve probabilistic scores in future studies.

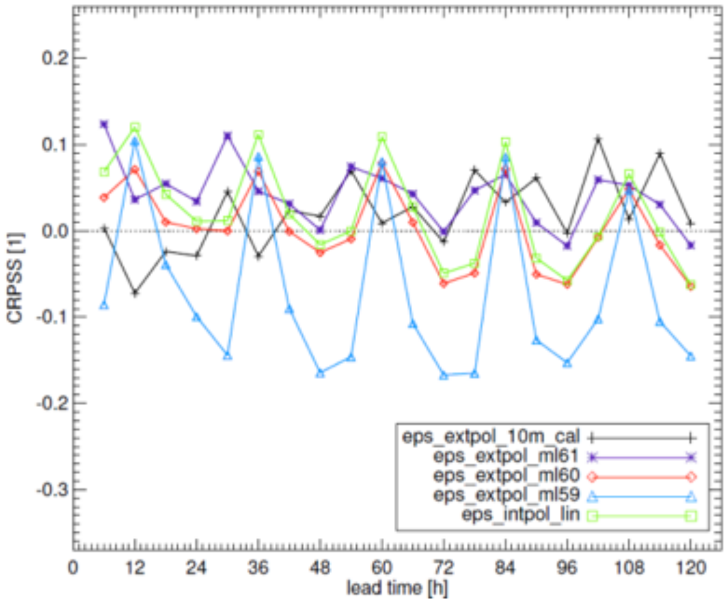


Figure 12: Improvement in skill (CRPSS) for an ensemble wind power prediction system for an EDF wind park utilizing ensemble winds from different heights: at 10-meters calibrated (black), model level 61 (35-meters, blue), model level 60 (67-meters, red), model level 59 (110-meters, turquoise) and linear interpolation between model level 61 and 59 (green). The reference ensemble system uses 10-meter uncalibrated ensemble winds. The time period is Jan-Oct 2008.

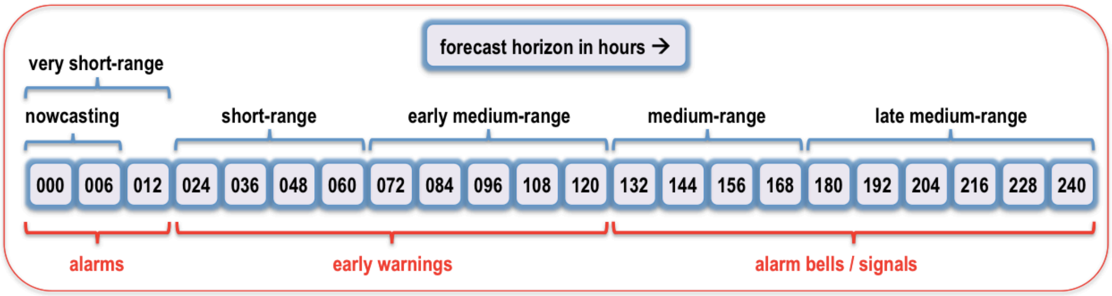
3. Utilisation of the ECMWF Extreme Forecast Index (EFI) – Set of examples

3.1 Introduction

The Extreme Forecast Index (EFI) was developed at ECMWF as a tool to provide forecasters with an indication of potential extreme weather events based on information from the ensemble predictions. Verification results [11] show that the EFI has substantial skill in forecasting extreme events several days in advance, confirming the subjective experience of forecasters in the Member States where the EFI is widely used. EFI skill is one of the six headline scores used to monitor long-term trends in performance of the ECMWF forecasting system [12].

The typical forecast horizon of EFI has been the early medium-range (3 to 5 days). During this time interval, EFI predictions of an extreme weather event can be considered as an "early indication". Beyond day 5, the EFI may serve as 'alarm bells' resulting from the ability of the ensemble to capture the risk of very intense weather systems (possible windstorms) at medium- and late medium-range. Box A contains a description of various terms used in this study: 'alarms', 'early indication' and 'alarm bells'.

Box A: Description of the graphical representation of different forecast and warning terms



'***Alarms***' refers to information concerning severe weather being anticipated in the very short-range. This type of information is based on methodologies or models capable of providing estimates about the level of predictability in the very short-term (mainly 0 to 6 hours while sometimes extending to 12 hours). Near-real time online observations are utilised in conjunction with immediate very short-term forecast updates on regional and local scales.

'Early warnings' refers to information about the occurrence of severe weather in the short- and early medium-term, i.e., in the next 12 to 60 hours (short-range) and 60 to 120 hours (early medium-range). Such tools, based mainly on the EFI, can be used to moderate risks and prepare users for effective response.

'***Alarm bells/signals***' refers to those cases for which very low probability extreme events can be captured by the IFS or some members (sometimes only one) of the EPS in the medium- or even in the late medium-range. As such signals become stronger and stronger, they should be considered as the basis (necessary elements) of issuing a more specific type of alert (i.e., an early warning).

Our investigation considers the process by which forecasters could make use of the EFI to extract information about future extreme weather events. The concepts are illustrated by studying the extreme winds affecting three airports in Germany. Results are presented for a synoptic study of extremes, skill assessment of the EFI and the possibility of setting optimal EFI thresholds for an early indication of windstorms. Finally some examples of utilising the EFI are given. It is intended that the results presented here will assist forecasters in providing warnings of high wind speeds.

3.2 Rare severe events

National Meteorological Services provide warnings about severe or high-impact events that can result in considerable damage and large losses. It is expected that much of the benefit to society through improved weather forecasts will come from advances in our capability to forecast such events so that mitigating actions can be taken. Indeed, one of the principal goals of ECMWF in the next ten years is to provide Member States' National Meteorological Services with reliable forecasts of severe weather across the medium-range while meeting Member States' requirements for high quality near-surface weather forecast products such as precipitation, wind and temperature.

Fortunately severe events tend to be rare events, hence the use of the term 'Rare Severe Event' (RSE) by *Murphy [13].* Such events are also loosely referred to as 'Extreme Events' in atmospheric science. RSEs can come in many forms, associated for example with very intense winds, heavy rain, extreme heat and cold, floods and droughts.

Forecasting RSEs poses specific problems because they are infrequent, poorly documented by observations, and at the limit of predictability. Quantitative verification of RSEs is therefore difficult and the statistical significance of verification results is mostly difficult to establish. At the same time, it is recognized that an imperfect numerical forecast in absolute terms can be of great value if it is well interpreted by an experienced forecaster. This means that a forecast error of given amplitude may have varying significance depending on where the forecast is placed with respect to the climatological distribution.

3.3 Predictability limitations concerning extremes

In operational forecasting, a 'gap' seems to exist between some of the events for which forecasters need to issue warnings and what the numerical model can provide guidance for. A study of past extreme wind events (such as windstorms) reveals that only a small proportion of ensemble members (or of single deterministic forecasts from different NWP centres) succeeded in predicting their true severity, even about 24 hours in advance. Some types of damaging or disruptive weather, such as lightning, wind gusts and fog, are not explicitly predicted by the models, and must therefore be inferred. Even if a type of weather can be explicitly predicted (e.g. heavy rain), the model resolution might be insufficient to capture its peak intensity; this could be because the associated processes are sub-grid scale. Several mesoscale models are being run experimentally at resolutions of 1–2 km, but most operational mesoscale models have grid scales of 5–15 km, and global models are even coarser.

Therefore we should not expect the current models always to reproduce the maximum values of weather parameters observed in extreme events because their resolution is relatively low. We should, however, design methods to diagnose severe weather based on the existing models, and thoroughly verify the validity of these diagnostics [14].

3.4 Extreme events and the EFI

The ability of models to generate extreme/severe storms with realistic frequency has improved significantly in recent years. Furthermore the development of ensemble prediction techniques has enabled the explicit representation of uncertainty in the forecast, both in the synoptic scale evolution and in the development of associated severe weather events. This means that models can now be used to provide information about the likelihood of extreme events occurring.

The Extreme Forecast Index (EFI) [15] has been developed to identify the risk of extreme events depending on location and season. The EFI measures the difference between the probability distribution of the ensemble forecast and that of the model climate. The underlying assumption is that if a forecast is extreme relative to the model climate, the real weather is also likely to be extreme compared to the real climate. The EFI is defined such that it lies between -1 and +1.

The EFI allows the forecaster to identify a possible future extreme weather situation without having to define specific thresholds for an extreme event. If the EFI indicates potential for a severe weather event, the forecaster can examine more detailed information from the forecast to make a more thorough assessment of the risk to the public.

Note that during the period covered by this study the resolution of the EPS has changed. Up to February 2006 it had a resolution of 80km, while up to January 2010 it had a resolution of 50km out to ten days – it then increased to ~30 km.

3.5 Dealing with extremes

Ensemble forecasts provide information on the uncertainty of forecasts. It is desirable to communicate this information particularly for events that can induce large losses. Probabilistic forecasts can also be used for decision-making by quantitatively assessing risk for specific users using a cost-loss model (for example). However, in the medium range, prediction of severe weather is likely to be associated with relatively low levels of confidence. Bearing this in mind, medium-range 'alarm bells' can ensure that potentially dangerous events do not go unnoticed by the forecasters.

In our investigation we consider events for which daily wind speed extremes exceed the 99th percentile of the model and station (synoptic) climate records. We will show that the EFI provides a useful indication of extreme events: high EFI values are generally associated with more extreme winds. By selecting an appropriate EFI threshold value, a user can tune their alert system to provide an optimal balance between hits and false alarms.

3.6 Case study for Bremen, Hamburg and Hannover airports

The link between extreme wind events and the EFI has been investigated for three synoptic stations based at airports in North Germany: Bremen, Hamburg and Hannover (as shown in Figure 13).

Figure 13: Geographical position of Bremen, Hamburg and Hannover airports/synoptic stations in North Germany (denoted by magenta dashed circles).

Two methods are used to define the wind speed extremes:

♦ *'Reanalysis' mode*. The ECMWF ERA-Interim [16] was used to construct a time series of daily maximum wind speeds for each station, spanning 2,374 days from 1 December 2003 to 31 May 2010. The maximum wind speed for each day was defined as the maximum value of the wind at the five synoptic hours: 00, 06, 12, 18 and 24 UTC.

♦ *'Observation' mode*. A time series was constructed based on each station's observations of maximum wind speed. In this case the daily maximum values are defined by considering 8 reported observations at 00, 03, 06, 09, 12, 15, 18 and 21 UTC.

The next step was to construct a time series of the daily maximum anomaly for each station in both 'Reanalysis' and 'Observation' modes. For each station and for all cases exceeding the 99th percentile, the synoptic meteorological environment was investigated. The extremes were linked to deep surface pressure lows, on most occasions affecting all three stations on the same day.

3.7 Detecting extreme events based on the EFI

The EFI is not only sensitive to a shift in the tails of the frequency distribution (i.e. in the extremes) but also to the median. In this study, the EFIs for two variables were utilised:

- ♦ *10FGI based on a maximum wind gust*
- ♦ *10WSI based on daily average of instantaneous 10-metre wind speed.*

For each of these, EFI forecasts based on both initialisation times (i.e. 00 and 12 UTC) were considered in 'Reanalysis' and 'Observation' modes.

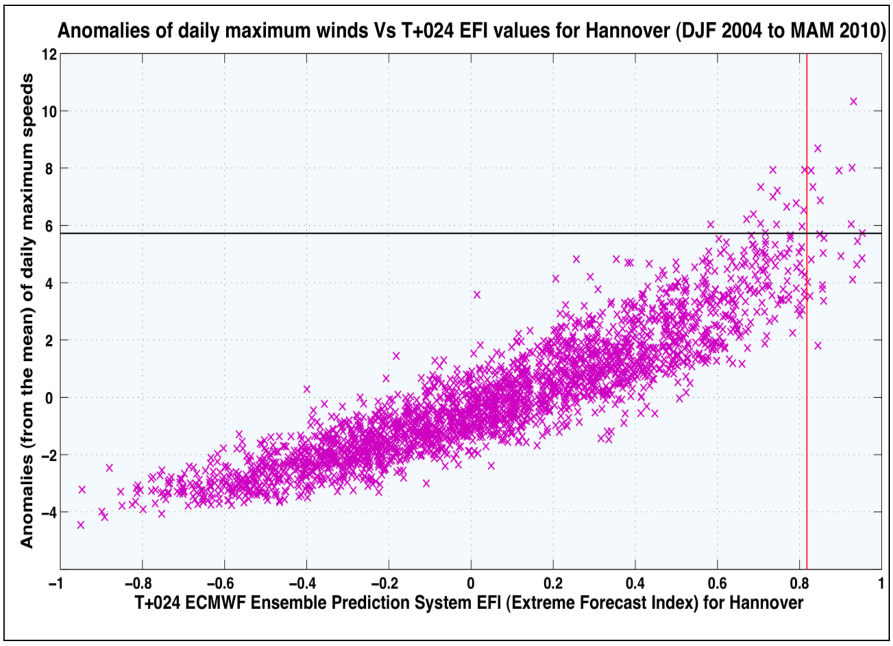


Figure 14: Example of anomalies of daily maximum 10-metre wind speeds in 'Reanalysis' mode against the 24-hour forecasts of EFI-10FGI (based on 00 UTC) values for Hannover. The red vertical lines represents the 99% EFI threshold, while the black horizontal line is the 99% percentile of maximum daily wind speed anomalies.

Clear signs that EFI values are closely linked to daily maximum wind speeds are contained in Figure 14. The 24-hour forecast is used in this example, but similar results apply for the other forecast horizons. These results reveal beyond any doubt that all reanalysis daily extremes (falling in the >99th percentile category) for Hannover correspond to strong positive EFI-10FGI values based on 00 UTC runs.

3.8 Skill assessment of the EFI

In addition to assessing point wise EFI values (Bremen, Hamburg and Hannover), their average wind maxima were also considered. Results in terms of hit rates and false alarm rates for different EFI thresholds are studied by utilizing ROC (Relative Operating Characteristic) diagrams and more specifically ROCA (Area under the ROC Curve) values. In terms of ROCA, the EFI-10FGI gust factors based on 00 UTC and 12 UTC data are comparable in skill in 'Reanalysis' mode, both comprising high values. Furthermore, the skill of the EFI forecasts over single points seems to be the same as that for the average of the three points. For EFI-10WSI no significant difference in skill was detected between forecasts based on 00 UTC and 12 UTC data in 'Reanalysis' mode. Also the skill of EFI-10WSI for selected points was found to be comparable to that obtained over the area covering Bremen, Hamburg and Hannover. Also in the 'Observation' mode there were no significant differences between using 00 and 12 UTC data in EFI-10FGI. The same applied to EFI-10WSI. However, for both the EFI-10FGI and EFI-10WSI the forecasts are found to be less skilful in 'Observation' mode. This is not surprising: the model wind (representative of a 50x50 km grid box) is not directly comparable to the observations at individual points. Another reason might be that the model has an easier task verifying against its own analysis (reanalysis for our case) extremes than against synoptic observations.

Overall EFI-10WSI was found to be less skilful as a forecast for maximum wind than EFI-10FGI. This could be anticipated since we constructed daily series of extreme wind values that are different from mean (daily averaged) wind time series in both 'Reanalysis' and 'Observation' modes. Going after such extremes, the EFI-10FGI formulation being based on model's 'gusty' components seems a more appropriate option than the EFI-10WSI formulation that is based on 'normal' instantaneous 10-metre wind components.

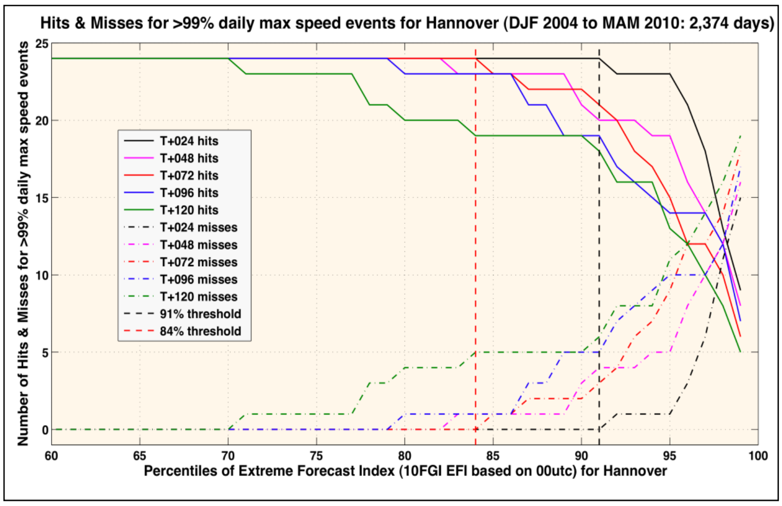


Figure 15: Hits and misses for the >99% category wind extremes based on different EFI-10FGI (00 UTC) thresholds for different forecast horizons (Hannover). The 91% EFI threshold (resulting to zero misses for day 1) and the 84% threshold (zero misses for day 3) are plotted as well (black and red vertical lines respectively).

Results in predicting extremes by utilising the EFI indicate significant skill in both the short- and early medium-range. It should be pointed out that to achieve high hit rates for all forecast horizons (as in the example shown in Figure 15), a significant number of false alarms would be generated as well. This behaviour is somewhat hidden by the rarity of the rare severe events represented in ROC curves and the associated ROCA scores [17]. However, early indications of potential extreme events allow users to take appropriate mitigating action. Depending on their sensitivity to the event, different users will take action at different levels of risk. A user who is especially vulnerable to an extreme event may decide to act even at a relatively low risk threshold, while others may prefer to wait until the event is more certain.

3.9 Setting an optimal EFI threshold

The usefulness of early indications of severe weather based on the EFI can be seen in Figure 16. This shows the EFI values for the maximum impact location (borders of Luxemburg and France) of storm Xynthia on 28 February 2010. It is clear that the EFI-10FGI is capable of providing an early indication of high winds four days in advance. The same holds for the other EFI variables but there is a delay of 24 hours.

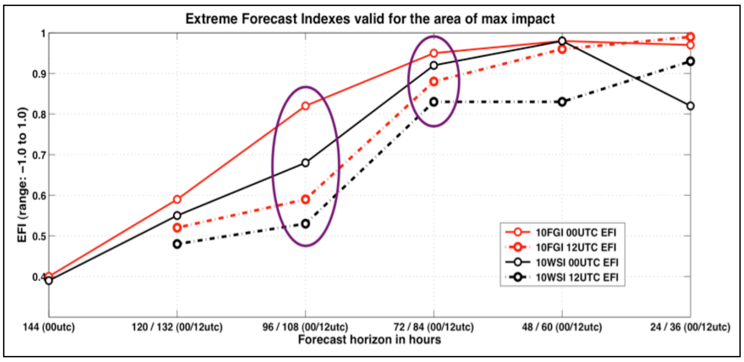


Figure 16: EFI-10FGI and EFI-WSI (based on 00 and 12 UTC) values for the area of Xynthia's maximum impact at the borders of Luxemburg and France (28 February 2010).

Using the 99th percentile of EFI, very high (skilful) ROCA values were found for all three airports. The 99-percentile threshold is capable of providing an early indication for some extremes, but not for all (as displayed in Figure 15).

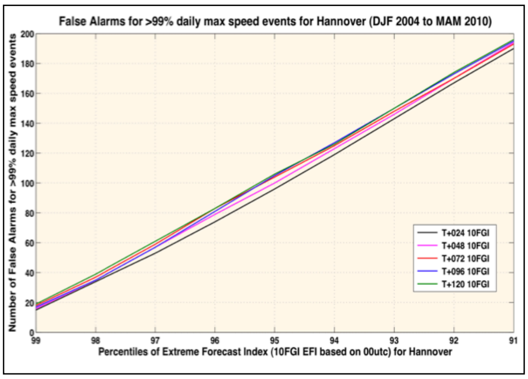


Figure 17: Number of False Alarms for different EFI percentile thresholds for horizons spanning from 24 to 120 hours. It is obvious that the 91% percentile that leads to zero misses for T+24 it also introduces 190 false alarms.

By lowering this threshold, the number of hits can be increased till eventually all extremes are captured, but the number of false alarms is then increased significantly. This unavoidable drawback can be seen in Figure 17 where the number of false alarms is plotted against different EFI-10FGI thresholds for Hannover airport corresponding to the hits contained in Table I.

Table I: Number of Hits for >99% category extremes based on various EFI-10FGI (00 UTC) thresholds for different forecast horizons valid for Hannover (maximum number of hits: 24).

EFI threshold (%)	Day 1 T+24	Day 2 T+48	Day 3 T+72	Day 4 T+96	Day 5 T+120
70	24	24	24	24	24
71	24	24	24	24	23
72	24	24	24	24	23
73	24	24	24	24	23
74	24	24	24	24	23
75	24	24	24	24	23
76	24	24	24	24	23
77	24	24	24	24	23
78	24	24	24	24	21
79	24	24	24	24	21
80	24	24	24	23	20
81	24	24	24	23	20
82	24	24	24	23	20
83	24	23	24	23	20
84	24	23	24	23	19
85	24	23	23	23	19
86	24	23	23	23	19
87	24	23	22	21	19
88	24	23	22	21	19
89	24	23	22	19	19
90	24	21	22	19	19
91	24	20	21	19	18
92	23	20	20	17	16
93	23	20	18	16	16
94	23	19	17	15	16
95	23	19	15	14	13
96	21	16	12	14	12
97	18	14	12	14	10
98	13	12	10	12	8
99	9	8	6	7	5

The number of hits for the 24-hour forecast is equal to 9, but there are also 15 misses and 15 false alarms (Table I). The 'zero misses' EFI threshold (i.e. the one corresponding to the 91st percentile), highlighted by yellow shading in Table I, is able to predict all 24 hits (i.e. zero misses), although by doing so the number of false alarms is increased significantly and reaches 190. This limitation becomes more pronounced when different (longer) time horizons are to be considered, as easily seen by examining the results for days 1 to 5 in Table I. For instance, the day 5 'zero misses' for the 99th percentile extreme wind anomalies correspond to a considerably lower threshold of EFI, equal to the 70th percentile (resulting in 688 false alarms). Overall, it is clear that all observed extremes (falling in the >99% category) are linked to high positive EFI values. The highest skill in providing an early indication is from the EFI-10FGI.

3.10 Examples of utilising the EFI

The setting of optimal EFI thresholds is further investigated for extreme events over Hannover. All daily maximum wind speed values for Hannover ('Reanalysis' mode) over a period of 2,374 days are plotted in Figure 18. A selection of the four most recent spikes has been made (highlighted by a red circle). These spikes indicate the following storms: Kyrill (18 January 2007), Emma (1 March 2008), Herbert (23 March 2009) and Xynthia (1 March 2010).

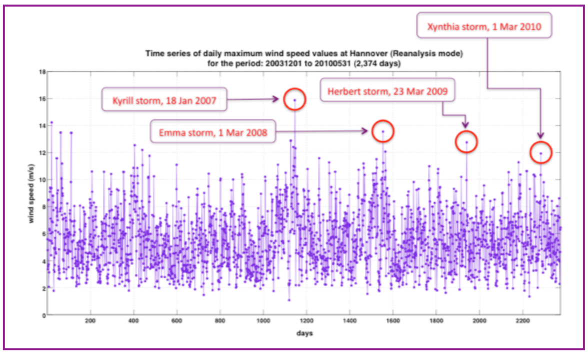


Figure 18: Time series of daily maximum wind speed values for Hannover over a period of 2,374 days from 1 December 2003 to 31 May 2010 ('Reanalysis' mode). Red cycles correspond to Kyrill, Emma, Herbert and Xynthia storms.

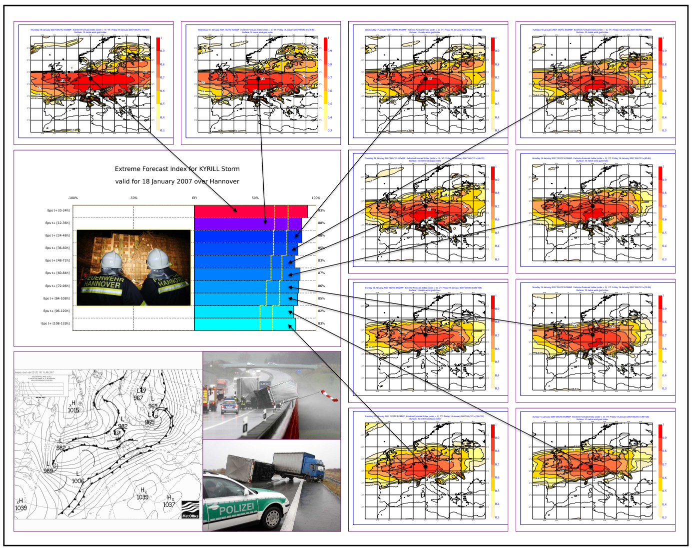


Figure 19: Example of different EFI (10FGI) maps ('EFI-GRAM') valid for the Kyrill storm hitting Hannover airport on 18 January 2007. The arrows from each map (initiating from Hannover's position) point to the part of the central graph constituting the currently operational 'EFI-GRAM' graphical set up for forecasters. The different forecast steps are displayed on the left of the gram while the exact EFI values over Hannover are displayed on the right. Forecast steps span from 24 to 132 hours with 12-hour intervals.

As an example the various EFI (10FGI) maps valid for Kyrill storm are displayed in Figure 19 for the forecast period from 24 to 132 hours with 12-hour intervals. It is clear that both the 95 and 98 percentile EFI thresholds (highlighted by a yellow dotted line) are able to provide an early indication of the Kyrill windstorm from day 5.5 (T+132 h) onwards, while Figure 20 corresponds to Emma storm.

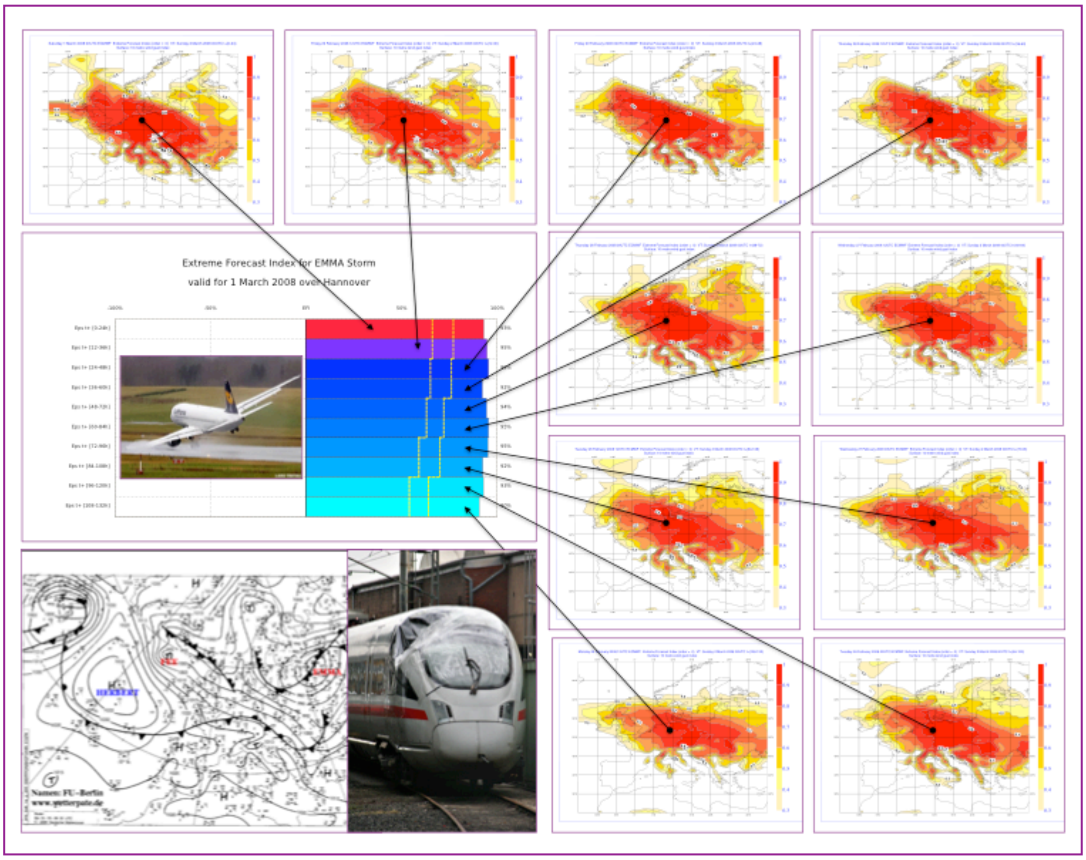


Figure 20: As in Figure 19, but for Emma storm hitting Hannover airport on 1 March 2008. The near crash incident of the Airbus A320 shown in the central picture took place at the nearby Hamburg airport.

In order to investigate whether these thresholds can provide an early indication of the other storms considered here, Figure 21 is constructed.

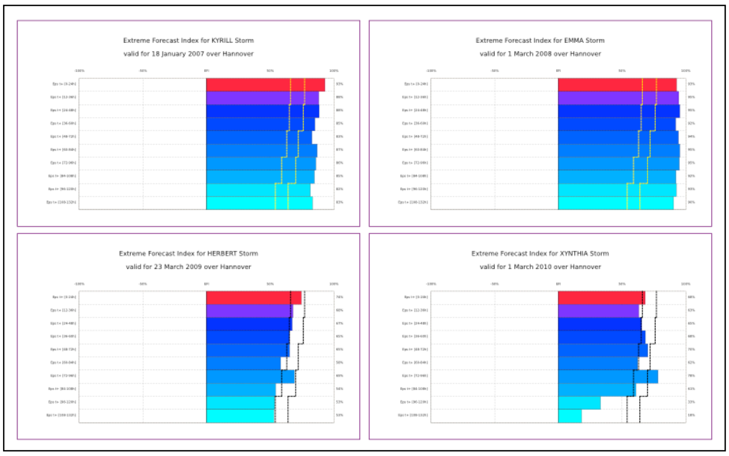


Figure 21: EFI values over Hannover for four windstorms (upper left panel) Kyrill, (upper right) Emma, (lower left) Herbert and (lower right) Xynthia. The 95% (left) and 98% (right line) thresholds are plotted in the upper panels using a yellow dotted line, while a black dotted line is used for the lower panes for reasons of clarity.

It is obvious that both the 95th and 98th percentile thresholds work quite well for the Kyrill and Emma storms, but they seem to be inadequate for Herbert and Xynthia. More specifically, for Herbert, using the 98th percentile threshold fails to give an indication of high winds, while use of the 95th percentile seems to do a better job for horizons shorter than 84 hours. As for Xynthia, the 98th percentile seems to work only for the 96-hour horizon, while the 95th percentile threshold works for all horizons shorter than 120 hours (except for the 36-hour one). For both Herbert and Xynthia, a slightly lower threshold (say a value between 90 and 95) could have resulted in forecasters having an early indication of the severity of the winds associated with the approaching storms.

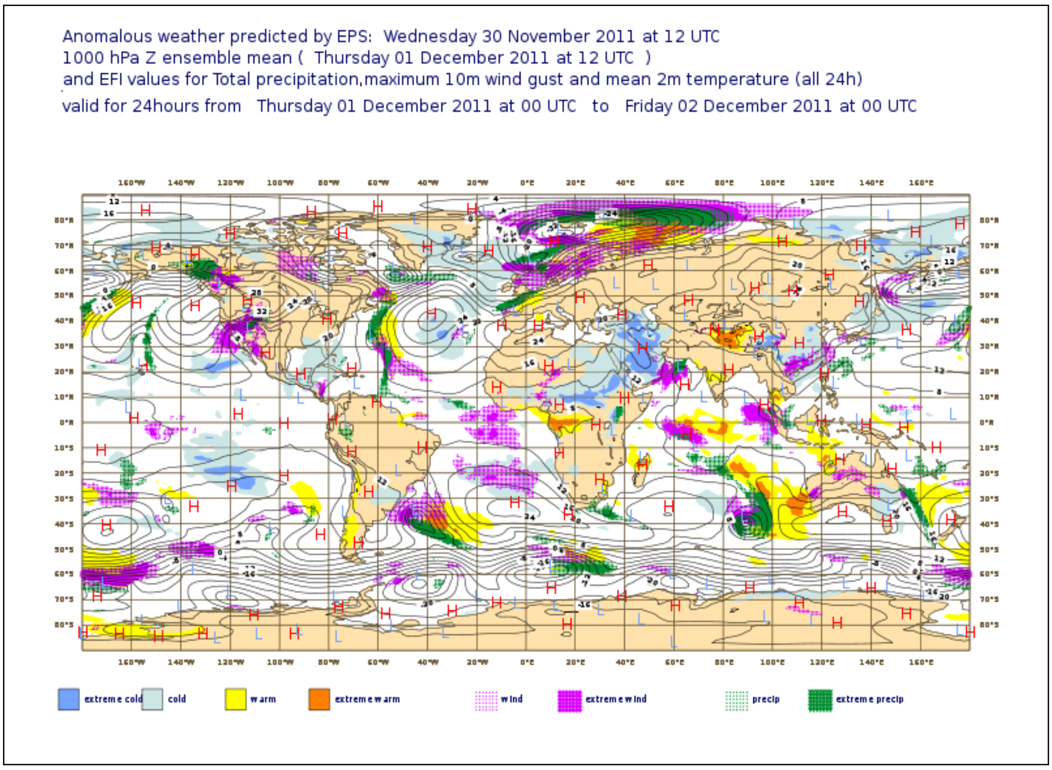


Figure 22: A composite I-EFI example, valid for 1 December 2011. Purple symbols correspond to possible wind extremes; green symbols correspond to rainfall, while the rest of coloured symbols correspond to temperature extremes.

An example of the operational global interactive EFI is given in Figure 22. By clicking and zooming over the area of interest the user gets all available information from EPS about the anticipated anomalous weather that might be linked to high-impact events.

4. Summary

The verification of CPS wind power forecast against real observed wind power in Germany arise the problem that wind speed forecasts in 10-meter are not optimal for wind power forecasting. In particular, as no information about the thermal stability of the atmosphere is available for the ensemble members, 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. CPS wind power forecasts have an improved forecast skill for extreme wind power penetrations. This is demonstrated for various German control zones and different penetration levels utilizing the Brier Skill Score. 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.

It has been demonstrated that the new EPS product of ECMWF to archive and disseminate winds at 100-meter height is a major achievement for probabilistic wind power forecasting. No wind power bias correction is necessary since the 100-meter winds can be directly interpolated to the hub height of modern wind turbines. This leads to an improvement of almost 50% in wind power RMSE (for Germany) using the 100-meter ensemble mean compared to the 10 m ensemble mean The superiority of 100-meter ensemble winds over 10-meter winds is also very striking in terms of probabilistic scores: The CRPSS for Germany is about 0.25 in the short-range and decrease in the early medium-range to a value of 0.1.

Furthermore, our investigation is focused on the early indication of extreme winds in the short- and early medium-range using the Extreme Forecast Index (EFI). For the assessment of the quality of the EFI, three synoptic stations at airports in North Germany (i.e. Bremen, Hamburg and Hannover) were considered. An investigation of synoptic weather type for each station indicated that all wind extremes (exceeding the 99th percentile) were linked to surface pressure lows being advected in south-westerly and north-westerly flow regimes. For the objective evaluation of early indications of an extreme weather event, the EFI for wind gusts and mean wind speed were compared to daily maximum wind speeds (in both ‘Reanalysis’ and ‘Observation’ modes). The highest skill in detecting extremes is given by the EFI-10FGI. ROCA (Area under the ROC Curve) values are found to be very high, indicating a skilful performance. Extreme observed events are clearly linked to higher values of the EFI.

Although the EFI is designed to be used qualitatively as a general “alarm bell” for potential extreme weather, it is also possible to use the EFI in a more quantitative way. The user can select a specific EFI threshold and take appropriate action whenever the EFI exceeds this threshold. The examples shown in this article illustrate some possible uses of this objective approach. There is no direct mathematical correspondence between percentiles of the EFI distribution and those of the climate distribution. However, in general selecting a high EFI threshold (e.g. the 99th percentile) focuses on the strongest warnings and will have fewest false alarms.

By lowering this threshold the number of hits is increased until all extremes are captured (i.e. zero misses), but by doing so the number of false alarms is increased significantly. Some users will be especially sensitive to missed events while others will be interested in limiting the number of false alarms. As our investigation has shown, each user is able to choose an appropriate EFI threshold for their own requirements, to provide an optimal trade-off between hits and false alarms.

5. References

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

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

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

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

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

[6] Deliverable 5.8 of the SafeWind Project, August 2012

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

[8] McLean JR (2008) Equivalent Wind Power Curves. Deliverable 2.4 of the TradeWind Project

[9] Deliverable 5.10 of the SafeWind Project

[10] Pinson P, 2012: Adaptive calibration of (u,v)-wind ensemble forecasts. Q.J.R. Meteorol. Soc., doi:10.1002/qj.1873.

[11] Richardson, D.S., J. Bidlot, L. Ferranti, A. Ghelli, T. Haiden, T. Hewson, M. Janousek, F. Prates & F. Vitart, 2011: Verification statistics and evaluations of ECMWF forecasts in 2010-2011. *ECMWF Technical Memorandum No. 654*.

[12] Andersson, E. & D. Richardson, 2011: Forecast performance 2011. *ECMWF Newsletter No. 130*, 15–16.

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

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

[15] Lalaurette, F., 2003: Early detection of abnormal weather using a probabilistic Extreme Forecast Index. *Q.J.R. Meteorol. Soc.*, 129, 3037–3057.

[16] Simmons, A., S. Uppala, D. Dee & S. Kobayashi, 2007: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter No. 110*, 25–35.

[17] Choo, C.W., 2009: Information use and early warning effectiveness: Perspectives and prospects. *J. Am. Soc Information Sci. Technol.*, 60, 1071–1082.