IEA Wind Recommended Practice for the Implementation of Renewable Forecasting Solutions: hands-on examples for the use of the guideline

Corinna Möhrlen[1] , John Zack[2], Mathias Blicher Bjerregård[3], Gregor Giebel[4]*

[1]*WEPROG, Assens, Denmark*
[2]*MESO Inc., Troy, NY, USA*
[3]*DTU Compute, Kgs. Lyngby, Denmark* [4]*DTU Wind, Risø Campus, Roskilde, Denmark*
**E-mail: com@weprog.com*

Keywords: IEA WIND, RECOMMENDED PRACTICE, RENEWABLE ENERGY FORECASTING, FORECAST EVALUATION, ENSEMBLE FORECASTING

Abstract

For any industry, it is important to establish standards and recommended best practices in order to ensure security of supply with a healthy competition structure. The IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions (IEA-RP) have been developed by a team of internationally active experts in wind and solar forecasting to fill this gap in the selection and implementation of optimal renewable energy forecasting solutions. The IEA-RP comprises four parts. The first part, Forecast Solution Selection Process, addresses the design of a customised process to select an optimal forecast solution for users specific situations. The second part, Design and Execution of Benchmarks and Trials, addresses the design, execution and analysis of customised forecasting benchmarks and trials. The third part, Forecast Solution Evaluation, describes methods and guidelines for meaningful evaluation of renewable energy forecasts and entire forecast solutions. The fourth part, Meteorological and Power Data Requirements for real-time Forecasting Applications, is a guideline for the selection, deployment and maintenance of meteorological sensors, power measurements and associated data quality control relevant to real-time forecasting. To assist in the practical application of the guideline, we provide three hands-on examples on how to use the guideline to design or improve forecast evaluation (Part 3) and measurement data quality (Part 4) in an efficient and impactful way. In the three use cases we demonstrate (1) evaluation of meteorological parameter forecasts (that could be used as input to a power prediction procedure) at a Danish coastal location, (2) verification of wind power predictions for a substation in the Northwest of Ireland and (3) quality control of meteorological measurements at an offshore location in the North Sea.

1 Introduction

The operational use of wind and solar power production forecasts has become widespread in the electric power industry and their benefits for the management of the variability of wind-based and solar-based generation have been documented in a number of studies (e.g., [1], [2]). However, while the operational use of forecasts has substantially grown over the past decade, there is considerable evidence that the full potential value of the wind and solar forecasts is often not realized in many applications. This is related to several factors. These include (1) specification of the wrong forecast objectives in the forecast solution selection process, (2) use of information from a poorly designed or executed forecast trial or benchmark to select a forecast solution, (3) use of forecast evaluation metrics that are not optimal for a user's application - that is they do not represent the way in which a user's application is sensitive to forecast error and (4) supplying meteorological or power production data from the generation facilities to the forecast process that is not of sufficient quality, representativeness or timeliness.

In order to address this issue, an international group of experts has worked under the structure of Tasks 36 and 51 of the International Energy Agency's (IEA) Wind Technology Collaboration Program (known as "IEA Wind") [3] to develop a set of four recommended practice documents (IEA-RP) that provides practical guidance on selection of optimal forecast solutions. The first phase of Task 36 extended from 2016 through 2018 and produced an initial version of the IEA-RP that consisted of three documents. The second phase of Task 36 was active from 2019 to 2021 and produced a second version of the IEA-RP that included revised versions of the original three documents and also added a fourth document that addressed issues with the gathering of data for input into the forecast process. The second version of the IEA-RP was also published by Elsevier as a book entitled as "The IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions" [4] The IEA-RP is designed to help streamline business processes for decision makers, system operators, traders, balance responsible parties and wind farm operators on a global basis.

The extension and refinement of the IEA-RP now continues under IEA Wind TCP Task 51. The fist phase of this Task began in January 2022 and will continue until the end of 2025. Information about the activities associated with Task 51 can be found at: `https://iea-wind.org/task51/`.

It is anticipated that the primary focus of the Task 51 work will be on the addition of datasets and tools to the IEA-RP document package. This will include a selection of use cases to serve as a resource for the industry in the adaptation and implementation of the recommendations. In addition as an assist to industry in the design and execution of forecast evaluation procedures, Task 51 is also developing an R-based forecast verification tool that facilitates the implementation of the forecast verification practices specified in the IEA-RP documents.

This paper provides practical guidance and real data examples of how to use the recommendations provided in the IEA-RP to evaluate alternative forecast solutions and ultimately select the best forecast solution for a specific application. The paper is organized in five sections. The section (Section 2) following this introduction provides an overview of the contents of the IEA-RP and the companion R-based verification tool. The next section (Section 3) presents an outline of how to effectively apply the IEA-RP guidelines in real world applications. Section 4 provides three specific use case examples. The paper concludes with a summary in Section 5.

2 Overview of the IEA-RP

The second version of the IEA-RP is composed of four parts. The first part, *Forecast Solution Selection Process* [5], addresses the design of a customised process to select an optimal forecast solution for user-specific situations. This is intended to provide guidance for the design of an economically viable process that will maximize the probability of obtaining an optimal forecast solution for a user's applications. Part 1 is divided into two core sections. The first is a discussion of the "big picture" issues that should be considered before starting the design of a selection procedure. The second is the presentation and discussion of a Decision Support Tool (DST) that steps through the issues that should be considered during the design of a forecast solution selection process.

The second part, *Design and Execution of Benchmarks and Trials* [6], addresses the design, execution and analysis of customised forecasting benchmarks and trials (B/T). For the purposes of the IEA-RP, a benchmark is defined as an exercise conducted to determine the features and quality of renewable energy forecast systems or methods such as those used to produce wind or solar power forecasts. A trial is an exercise conducted to test the features and quality of operational renewable energy forecast solutions. Part 2 provides guidance for optimizing the three fundamental phases of a B/T: (1) preparation, (2) execution, and (3) evaluation and decision-making. Optimizing the design specifications and execution protocols in each phase dramatically increases the probability that the B/T will provide meaningful data to support the selection of an optimal forecast solution for a particular application.

The third part, *Forecast Solution Evaluation* [7], describes methods and guidelines for meaningful evaluation of renewable energy forecasts and entire forecast solutions. The evaluation process is a large component of the forecast solution selection process if a benchmark or trial is conducted as part of the process but an evaluation is also an important component of an ongoing performance assessment program. Part 3 provides guidance for the effective evaluation of the performance of alternative forecast solutions. The guidance is based on four fundamental principles: (1) evaluation is subjective, i.e. it is important to understand the limitations of chosen metrics, (2) evaluation has an inherent uncertainty due to its dependence on the evaluation dataset and the specific metrics that are employed, (3) evaluation should contain a set of metrics in order to measure a range of forecast performance attributes and (4) evaluation should reflect a "cost function", i.e. the selected metric combinations should provide an estimate of the value of the solution for the specific target applications(s).

The fourth part, *Meteorological and Power Data Requirements for real-time Forecasting Applications* [8], is a guideline for the selection, deployment and maintenance of meteorological sensors, power measurements and associated data quality control relevant to real-time forecasting. The focus is on the impact that measurement-data-related decisions that affect the characteristics (e.g. availability, quality, representativeness, timeliness, etc.) of the data available from a wind or solar generation facility ultimately have on wind or solar generation forecast performance.

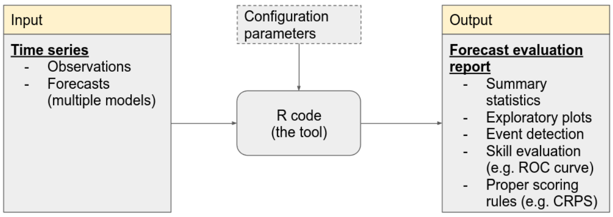

Fig. 1 Concept of the Companion Forecast Evaluation Tool (*WE-verify-prob*) to the Recommended Practice part 3 [9]

In order to provide practical support for the guidance provided in these four documents, a companion forecast verification tool called *WE-verify-prob*[10] is being developed to facilitate the design and execution of holistic evaluations of forecast performance. WE-verify-prob[10] is an R-based code base to verify probabilistic wind energy forecasts. It provides example code for many of the probabilistic forecast evaluation metrics described in the IEA-RP Part 2 Designing and Executing Forecasting Benchmarks and Trials [11] and Part 3 Forecast Solution Evaluation [9]. The tool is accessible for download at the IEA Wind Task 51 web page* or as an R-package. The development of the tool is expected to continue under Task 51. The use of the term "holistic" refers to the assessment of a wide range of forecast performance attributes, which is in contrast to the dominant forecast user practice of assessing only the "typical forecast error" attribute as evaluated by popular metrics such as the "mean absolute error" or the "root mean square error". The conceptual design of the tool is illustrated in Fig. 1.

*`https://iea-wind.org/task51/`

The tool accepts two types of input: (1) time series of observed and forecasted data and (2) a set of configuration parameters. Once the user has supplied the required inputs, the tool generates reports containing a set of forecast verification statistics and explanatory plots.

3 Practical Use of the IEA-RP Recommendations

This section provides an outline of the how to use the recommendations in the IEA-RP in real world applications. Although many of the principles and guidelines specified the IEA-RP are likely to be useful in a broad range of forecasting applications, they are specifically intended for the following application areas (see [12]):

i) System Operation, Balancing and Trading:

- Situational awareness in critical weather events
- High-Speed Shutdown events
- Grid related down-regulation or curtailments
- Short-term forecasting with updates from measurements
- Intra-day power plant balancing

ii) Wind Turbine, Wind Farm and Solar Plant Operation and Monitoring:

- Wind turbine and Power Plant Control
- Condition Monitoring

3.1 Selection of Instrumentation

The recommendations for the selection of instrumentation to provide data for input into the forecast process is based on the consideration of accuracy and reliability requirements. Accuracy requirements need to be defined for the application/project and aligned with the associated levels of effort necessary to operate and maintain the measurement system under these constraints. An overall cost-performance determination should therefore always be carried out to adapt the budget to the accuracy requirements and vice versa. Reliability can be achieved with redundant instrumentation and/or high quality instrumentation. Redundancy enhances and ensures confidence in data quality. Selection of multiple instruments need to be aligned with the accuracy needs.

3.2 Gathering a Meaningful Forecast Evaluation Sample

A key component of a forecast evaluation is the gathering of a meaningful forecast evaluation sample. "Meaningful" in this context is defined as a sample of cases that is representative of the range of forecast scenarios that are likely to be encountered during a typical year at the target sites. Such a sample can be ideally constructed by obtaining forecasts and the corresponding observational data for one or more years. In most cases this is not feasible and it is necessary to explicitly construct a sample that includes a sufficient representation of key forecast scenarios which should include seasonal variations (e.g. high resource (wind, solar) season, low resource season and typical types of extreme (i.e. the tails of the resource distribution) or difficult to forecast events. If a sample is to be gathered in real-time then the exercise should be run in several periods at different times of the year. If the sample is gathered from historical data, then the key types of scenarios should be identified and the sample should be constructed to include representative cases of each key type of forecast situations.

3.3 Specification of Appropriate Forecast Evaluation Metrics

Once a meaningful forecast sample has been gathered, the next step is to construct an appropriate set of forecast performance metrics for the target application. As discussed in the IEA-RP (see section 15.1.5. of [9]) it is desirable to use a set of metrics to measures a range of aspects of forecast performance rather than a single metric or a small set of metrics that measure one attribute such as the "typical error". The set of metrics should represent or approximate the sensitivity of the target application to forecast error. Ideally, this should take the form of a "cost function" that quantitatively links the cost of a forecast error to the errors in the key forecast variables. However, in many cases a true "cost function" is not available for a particular application. A reasonable alternative is then to use a set of metrics that approximate a user's understanding of how the application is sensitive to forecast error. For example, are errors more critical in certain scenarios or times of the day or year, different forecast lead-times or threshold limits? Or are errors that have a magnitude below a specific threshold not important? (i.e. are errors only significant to the application, if they are outside of a specified range of noise)?

Another aspect to consider is “reducibility” and “stability” of results. For example the verification of probabilistic forecasts with the reliability diagram has previously caused concern in the community, because the choice of bins in the classical approach with equidistant bins is prone to generate drastically distinct reliability diagrams that easily can give a wrong impression of the goodness of a forecasting system - to both sites. In [13] it was reported that “..alternative methods for the choice of the binning have been proposed in the literature, extant approaches exhibit similar instabilities, lack theoretical justification, are elaborate, and have not been adopted by practitioners”. Another way of verifying the reliability of a probabilistic forecast is to use a continuous binning approach, the so-called “Consistent, Optimally binned, and Reproducible reliability diagrams using the Pool-adjacent-violators algorithm” (CORP) approach (see [13]).

The CORP approach is reported to resolve these issues in a theoretically optimal and readily implementable way for practical verification tasks for reliability diagrams and score decompositions and hence has been implemented into our example verification code “WE-verify-prob” [10].

3.4 Production and Interpretation of Forecast Evaluation Metric Data

While it is conceptually straightforward, the production and interpretation of forecast evaluation metric data can pose some significant issues to inexperienced evaluators and can also be a significant source of noise in the final set of performance metric data. Evaluators with an experience with many benchmarks and trials are often amazed at the difference in forecast metric data that arises when the same set of forecast metrics are computed with the same set of input data by different evaluators or software modules. There are many overlooked factors that contribute to these differences including how missing forecast or observed values are treated, how the data is quality controlled (i.e. which input data are classified as unacceptable) and the computational procedure (formulas, order of operations etc.) used to compute the metrics. These issues typically become more significant in the evaluation of probabilistic forecasts since the forecasts themselves are more complex and therefore the formulation of the metrics is more complex. This makes them more challenging to compute - especially in a spreadsheet environment - and also to interpret. In order to address this issue the *WE-verify-prob*[10] software is being developed (see also section 2).

4 Example Use-Cases for the Recommended Practice Guideline

4.1 Wind speed Comparison at a Danish SYNOP station of a test-setup with a running setup

The first use case example is a comparison of two alternate forecasts of wind speed for a wind measurement site. The location is a synoptic meteorological observing station (SYNOP) that reports 10-m wind speed and direction at a complex coastal location near Assens, Denmark (55°14’50.3"N 9°53’24.8"E | 55.2473N, 9.89023E). For this comparison two 10-m wind speed forecasts from a 75-member forecast model ensemble based on WEPROG’s Multi-Scheme Ensemble Prediction System (MSEPS) were used: (1) a high resolution (5 km grid cells with 60 vertical levels) forecast modeling system labeled "HR" and (2) a lower resolution (15 km grid cells with 32 vertical levels) operational modeling system labeled "LR".

The WE-verify-prob software [10] (see also 3.4) was used to assess the performance of these two sets of forecasts. We start our comparison evaluation with the CRPS score [14] for a 6-11 hour forecast and a 0-48 hour forecast, shown in Table 1.

Additionally, we decompose the score in lead-time dependent score values for each hour of the forecast horizon to evaluate the error growth of the two forecast systems, shown in Fig 2.

The CRPS score is the equivalent of the mean absolute error (MAE) score for probabilistic forecast and can also be interpreted as the integral of the Brier score over all possible threshold values for the parameter under consideration. As with the MAE, the lower values of the CRPS indicate better performance. For statistically generated probabilistic forecasts (e.g. with quantile regression), the cumulative distribution function (CDF) is known and can be used to compute the CRPS. For numerical ensemble weather prediction models with different model physics and/or initial conditions, it is necessary to convert the data into an estimated (cumulative) distribution function that can be evaluated at any point z ∈ R [15]. As Zamo and Naveau [16] pointed out, the CRPS is estimated with some error, when the true forecast CDF is not fully known, but represented as an ensemble of values and that “..using the CRPS to compare parametric probabilistic forecasts with ensemble forecasts may be misleading due to the unknown error of the estimated CDF for the ensemble”. In our case, we compare two forecasts generated with the same ensemble prediction system in different setups and hence we can ignore this potential error source.

Table 1 CRPS for a 48 hour ahead forecast for the high-resolution (HR), low-resolution (LR) ensemble, the reference and the improvement relative to the reference.

Forecast Type	CRPS	Improvement to Reference [%]
Reference	1.6635	
Lead-time	**6-11h**	
HR	1.140	-31.5
LR	1.159	-30.3
Lead-time	**0-48h**	
HR	1.1236	-32.5
LR	1.0925	-34.3

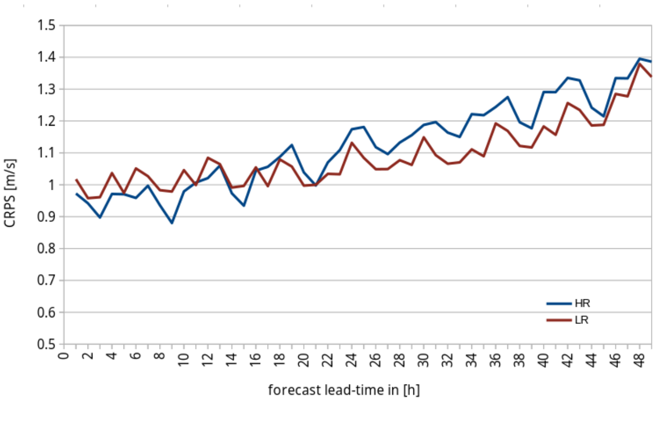

Fig. 2 CRPS by lead time for the high-resoluion (HR) and the low-resolution (LR) setup, showing the error growth over 48 hours.

The high-resolution forecast HR scores better than the operational low-resolution LR forecast for the coastal site in the first 12 forecast hours. This is an expected result, especially for a coastal site, where the higher resolution is resolving the coastline much better than a coarser resolution model setup. Both forecasts are approximately 30% better performance in comparison to the reference forecast and the high-resolution forecasts improves 1.5% over the low-resolution operational forecast. However, when looking at lead-times above 12 hours, the LR forecast has a lower error growth in comparison to the HR forecast, which can be seen in Figure 2 and the CRPS for the 0-48 hour lead-time in the second part of Table 1. The non-uniform lines are due to the changing forecast initial times. Nevertheless, our example demonstrates well, that it is important to test and verify forecasts on the exact target in order to generate a fair verification, but also to take the right decision from the result. The slightly higher error growth for the high resolution system over the longer forecast lengths are due the larger amount of degrees of freedom in the higher resolution and the fact that the LR forecast is tuned for good average scores in the day-ahead forecast horizon.

Additionally, the reliability of the forecasts was evaluated. Reliability is the degree to which the forecasted probabilities are in agreement with the outcome frequencies. This evaluation is best done by constructing a reliability diagram. The reliability diagrams are constructed with the R-package “reliabilitydiag: Reliability Diagrams Using Isotonic Regression” [17], which is also part of the WE-verify-prob verification-tool[10]. The package checks the reliability of predictions with the so-called “Consistent, Optimally binned, and Reproducible reliability diagrams using the Pool-adjacent-violators algorithm” (CORP) approach [13] (see more detailed description in section 3.3).

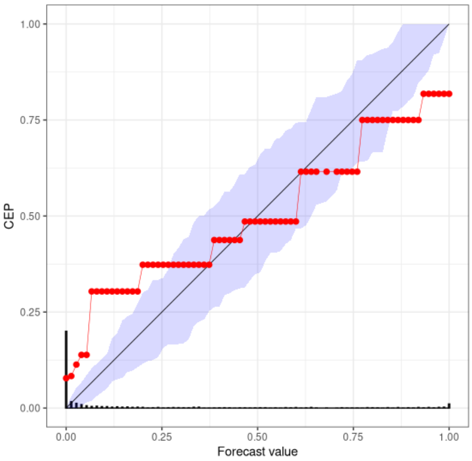

Fig. 3 Reliability Diagram and histogram for Forecast HR (high resolution) of the 10-m wind speed at a meteorological station on the coast of Denmark with the conditional event probabilities (CEPs) on the y-axis and respective probabilities at the x-axis. The blue consistency band is the 90% uncertainty quantification.

Those not so experienced with the reliability diagram and its features, can seek help in Figure 14.5 of the IEA-RP [18] or [19] which illustrate various examples of reliability diagrams and how information about calibration , resolution and uncertainty can be inferred from the graphs.

For our example we have use a threshold of 5, which means that a minimum of 5 “positives” are needed for an event, a

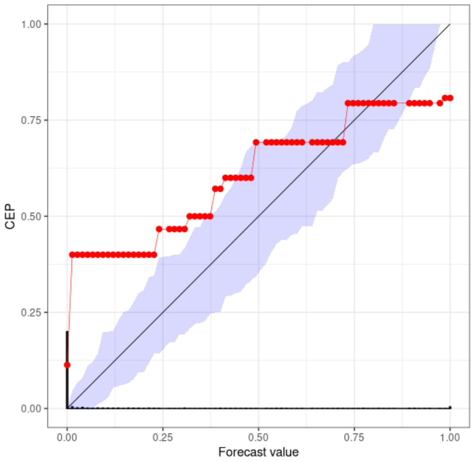

Fig. 4 Reliability Diagram for Forecast LR (low resolution) of the 10-m wind speed at a meteorological station on the coast of Denmark with the conditional event probabilities (CEPs) on the y-axis and the respective probabilities at the x-axis.The blue consistency band is the 90% uncertainty quantification.

forecast horizon of 6-11 hours with a change in wind speed of 3m/s over a 3 hour window. Note that the choice of threshold and event definition is rather sensitive. It is therefore recommended to be very careful in the choice of these limits and threshold values for the event definitions and choose these carefully according to the desired target outcome. We can see in the examples in Fig. 3 and Fig. 4 that the low-resolution LR setup shows a tendency to lie on top of the diagonal, except for the very upper probability bins. This indicates a negative BIAS and/or a slight mis-calibration. In the lower probability bins, there are in both cases somewhat extended horizontal segments that indicate additionally a diminished discrimination ability. For the HR setup, this is only pronounced for those parts within the lower probability bins. Otherwise the HR setup has a better balance between resolution and calibration, staying mostly within the blue 90% consistency band. The fact that both indicators for lack of resolution and calibration are more pronounced for the low-resolution setup, they confirm the expectation of the high-resolution setup being superior to provide a reliable probabilistic wind speed forecast with a reasonable resolution. Overall we can conclude from the evaluation that the high-resolution system has a better representation of the uncertainty in the ensemble and a higher accuracy. The next step would be a cost-benefit analysis, which often involves an analysis of the improvements for other end products such as wind and solar forecasts. The next use-case will deal with that.

4.2 Probabilistic Wind Power Forecast Comparison

The second use case example illustrates the use of a set of metrics to compare the performance of two probabilistic forecasts from different forecast model configurations for the prediction of the wind power time series and the occurrence of wind ramp events of a specified magnitude and time duration. The objective is to select one of the configurations for operational use based upon the diagnosed differences in forecast performance. In this example the two alternative forecasts are from the same two setups of the WEPROG's MSEPS 75-member ensemble system as in use case 1 (see section 4.1, but this time converted into wind power forecasts. They are also labelled as "HR" for the high resolution and "LR" for the low resolution of the same system. The forecasts are for the wind power generation that is fed into a sub-station in the Northwest of Ireland. There are a few wind generation plants connected to this substation and the aggregate capacity is 180 MW. The test period extended for 3 months from mid February 2023 to mid May 2023. The evaluation will focus on the the 6- to 11-hour look ahead period. Once again the WE-verify-prob software [10] was used to verify power generation predictions from the two sets of forecasts.

The first attribute to be evaluated is the "typical" error of the probabilistic 1-hour wind power forecasts during the 6 to 11 hour look-ahead period. The CRPS is useful for this objective. As noted previously the CRPS is the probabilistic analogue to the MAE for a deterministic forecast. Lower CRPS values indicate smaller error and therefore better performance. The CRPS scores for each forecast over the 3-month test period are listed in Table 2. These results indicate that both forecasts have a much better score than a reference probabilistic forecast and that the HR configuration performs slightly better than the LR configuration. The reference forecast here is a distribution constructed from all observations in the time series and can be compared to a persistence forecast for deterministic verification, where the observed value at present time will persist into the future. The Glossary of Meteorology from the American Society of Meteorology describes the persistence forecast as often being "..used as a standard of comparison in measuring the degree of skill of forecasts prepared by other methods, especially for very short projections".

Table 2 CRPS with results given in MW and percent of installed capacity.

Forecast	CRPS	CRPS
	[MW]	[% inst. cap]
HR	10.5	5.8
LR	10.9	6.0
Reference	20.6	11.5

Next, the focus of the evaluation procedure shifts to the assessment of the ability to predict wind ramp events during the forecast window. Four ramp event thresholds will be considered: (1) 20 MW change in hour, (2) 30 MW change in 3 hours, (3) 40 MW change in 3 hours and (4) a 60 MW change. The Brier Score (BS) is the most commonly used metric to assess the overall accuracy of a probabilistic event forecast. It measures the mean squared difference between the forecasted probability (e.g., 0 to 1) and the actual outcome (e.g., 0 or 1). Thus, it is analogous to the mean squared error of a deterministic forecast. The BS values range between 0 and 1 with lower values indicating better performance. A value of 0.0 is a perfect forecast which can only be achieve if a forecasted probability of 1.0 is made for every event that occurs and a probability of 0.0 is predicted for every non-event. The BS for each wind ramp threshold for the two sets of forecasts is shown in Table 3. The bottom row of the table also shows the the difference in BS between the HR and LR forecasts. The BS metric indicates that the LR forecasts slightly outperform (i.e. lower BS) the HR forecasts for all of the thresholds. Often is it said that the rarer an event, the easier it is to get a good BS without having any real skill. This is also the case here, where the difference between the HR and the LR is least for the 60MW over 3 hours ramp in comparison to the other ramping limits.

Table 3 BRIER SCORE for different ramping limits and time window.

Forecast	20MW 1hour	30MW 3 hours	40MW 3 hours	60MW 3 hours
HR	0.0501	0.089	0.0513	0.021
LR	0.0459	0.084	0.0464	0.018
$\Delta(HR-LR)$	0.0043	0.0053	0.0049	0.0028

Additional information about the forecast performance can be obtained by decomposing the Brier score into three components: (1) calibration/reliability (CAL), (2) discrimination/resolution (DSC/RES) and uncertainty (UNC). The CAL term measures the degree to which the forecasted probability agrees with the frequency of event occurrence given the forecasted probability (conditional event probability). This attribute is often referred to as the reliability. In the decomposition of a perfectly reliable forecast (i.e. the frequency of occurrence always matches the forecasted probability) a CAL score would have a value of 0.0. In this case, it is often said that the forecast is "well-calibrated". The second term (DSC/RES) measures the ability of the forecasts to correctly distinguish differences in the probabilities among the cases. This term has a negative sign in the decomposition equation so higher values contribute to lower BS values and therefore indicate better performance. The UNC term measures the inherent uncertainty in the event and is related to the event frequency in the evaluation sample. Lower values of this term contribute to lower BS values. The maximum UNC occurs when a event occurs in half of the cases and the minimum UNC occurs when the event always occurs or never occurs. Note that this term does not depend on the forecast and indicates that the BS will be higher for more uncertain events regardless of the forecast performance.

The values of the three BS decomposition terms for the two forecasts and the four ramp event thresholds are shown in Table 4, which lists the mean score (as shown in Table 3) and the values of the three decomposition terms. The CORP approach[13] was employed to compute the BRIER decomposition for this example. The decomposition terms are semantically the same as the description above and in chapter 14.3.1 in [18], but mathematically slightly different and statistically more proper. The main difference is that the classical components from Murphy's decomposition[20] are, according to Dimitriadis et al. [13], only exact in the discrete case, but fail to be exact under continuous forecasts, which we use here. The mathematically modified components all refer to a reference forecast, which is defined as the marginal event frequency.

The results in Table 4 indicate that the LR forecasts exhibit better reliability (lower CAL score) and higher resolution (higher DSC/RES score) for all four ramp thresholds. The UNC score is the same for both forecasts, because as noted previously, it depends only on the frequencies of the events in the sample and not on the forecast attributes. The 30MW/3hr events are most frequent in the sample and therefore have the highest UNC score. It should be noted that the mean BS is highest for this event for both forecasts due to the high value of UNC.

Table 4 BRIER's score de-compostion for different ramping rates and time windows. The de-composition provide important additional information to the reliability diagrams with the mean-score (MS), (mis-)calibration (CAL), discrimination (DSC/RES) and uncertainty (UNC).

Fore- cast	MS	CAL	DSC (RES)	UNC
Limit:	30MW/3h			
HR	0.0892	0.0105	0.0274	0.106
LR	0.0839	0.0062	0.0283	0.106
Limit:	40MW/3h			
HR	0.0513	0.0074	0.0153	0.0592
LR	0.0464	0.0029	0.0157	0.0592
Limit:	60MW/3h			
HR	0.0210	0.0018	0.0024	0.0217
LR	0.0182	0.0010	0.0045	0.0217
Limit:	20MW/1h			
HR	0.0501	0.00494	0.00457	0.0498
LR	0.0459	0.00248	0.00639	0.0498

The reliability of the forecasts can also be viewed from a graphical perspective. This is typically done by plotting the forecasted probabilities on the x-axis and the corresponding conditional event probabilities (CEP) (i.e. the frequency of observed events given the specific forecast probability) on the y-axis. Reliability diagrams for the HR and LR forecasts of the 30 MW/3hr ramp event threshold are shown in Figures 5 and 6. The red markers represent the histogram of the actual values from the forecast evaluation and the closer these are to the diagonal "perfect agreement" line, the higher the forecast reliability. The diagrams qualitatively indicate that the LR forecast is more reliable than the HR forecast, which is consistent with the CAL scores shown in the first two rows of Table 4. However, the diagram provides more insight into the nature of the reliability issues with the two forecasts. For example the probabilities of the HR forecasts tend to be too low especially for low forecasted values, whereas the LR forecasts tend to be slightly too high except for low forecasted values. In other words, the LR setup lacks some resolution, but stays more within the 90% consistency band, while the HR has a negative BIAS (with most values above the diagonal) and more values outside the 90% consistency band.

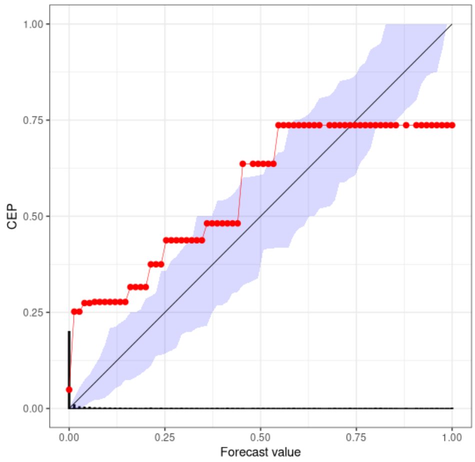

Fig. 5 Reliability Diagram and histogram for the high-resolution (HR) forecasts of the 30 MW/3hr wind ramp events with the conditional event probabilities (CEPs) on the y-axis and respective probabilities at the x-axis. The blue consistency band is the 90% uncertainty quantification.

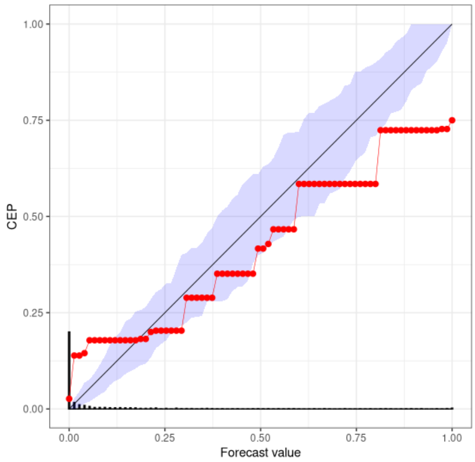

Fig. 6 Reliability Diagram and histogram for the low-resolution (LR) forecasts of the 30 MW/3hr wind ramp events with the conditional event probabilities (CEPs) on the y-axis and the respective probabilities at the x-axis.The blue consistency band is the 90% uncertainty quantification.

It often useful to examine the event forecast performance from a deterministic perspective, even if the underlying forecast is based on a probabilistic forecast approach. This is typically done by either setting a threshold probability (presumably after calibration) to produce a binary event/no-event forecast or by counting the number of ensemble members to determine whether the majority (or some other counting threshold) produce either events or no events. Once a set of deterministic forecasts are derived than a contingency table approach can be used to assess the performance of the forecasts. A contingency table was generated by the WE-verify-prob tool for the four wind ramp thresholds considered in this example. The conversion from the 75-member ensemble data set to the binary event forecast was performed with an algorithm within WE-verify-prob. The resulting contingency table is shown in Table 5. This table lists the absolute number of "hits", "misses", "false alarms" and "correct negatives" in the forecast sample and also the "hit rate" (HiR) which is the hits per total number of forecasts and the "false alarm rate", which is the false alarms per total number of forecasts.

The results indicate that the LR forecasts have a much higher hit rate (HiR) for all thresholds but also have a somewhat higher FAR for each threshold. The most extreme example of this pattern is for the 60MW/3hr threshold. The LR forecasts have approximately a three times higher HiR, but also a more than 3 times higher FAR.

Table 5 Contingency table inclusive hit rate (HiR), false alarm rate (FAR).

Fore- cast	Hits	Misses	False Alarms	Correct Neg.	HR	FAR
Limit:	**30MW**	**window:**	**3h**			
HR	149	145	153	1990	0.507	0.071
LR	204	90	393	1750	0.694	0.183
Limit:	**40MW**	**window:**	**3h**			
HR	82	72	91	2192	0.532	0.04
LR	112	42	262	2021	0.727	0.115
Limit:	**60MW**	**window:**	**3h**			
HR	10	44	31	2352	0.185	0.013
LR	30	24	102	2281	0.556	0.043
Limit:	**20MW**	**window:**	**1h**			
HR	37	91	101	2208	0.289	0.044
LR	74	54	302	2007	0.578	0.131

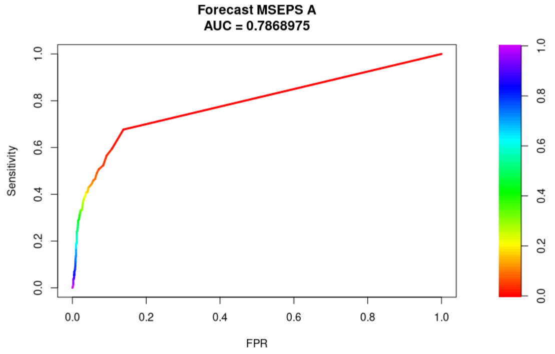

Fig. 7 ROC Curve for the 30MW/3hr ramp event threshold for the high-resolution (HR) MSEPS

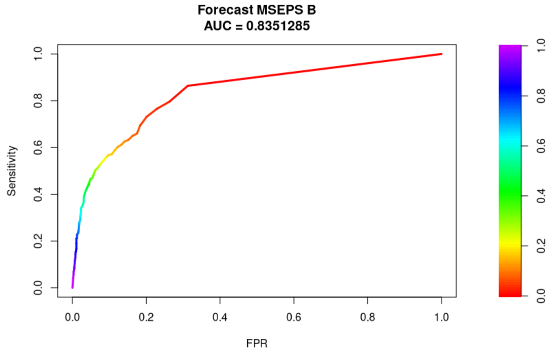

Fig. 8 ROC Curve for the 30MW/3hr ramp event threshold for the low-resolution (LR) MSEPS.

Another useful tool to measure the ability of a forecast to deterministically discriminate between events and non-events is the Receiver Operatering Characteristic (ROC) curve. This is a plot of the FPR ("false positive rate") vs the HiR (hit rate also known as the "sensitivity") for all classification thresholds. In this case, the classification thresholds are different values of the ramp rate value used to separate forecasted ramp rates into the "event" or "no event" categories. ROC curves for the HR and LR forecasts of the 30MW change in 3 hours ramp event are shown in Figs 7 and 8. Visually, it is noticeable that the ROC curve for the LR forecasts goes further to the top of the chart sooner, when going from left (low FPR) to right (high FPR) on the chart. This indicates that the sensitivity (hit rate) is higher for a given FPR in this range of the chart.

The "area under the curve" (AUC) is used to summarise the overall accuracy of the event forecast (see Table 6). Ideally, the curve ascends vertically at FAR=0.0 and goes horizontally at a sensitivity (hit rate) value of 1.0. This means every forecast is a hit and there are no false alarms regardless of the classification threshold that is selected. In this case the AUC is 1.0, which is a perfectly accurate forecast. The theoretically lowest value of 0 indicates a perfectly inaccurate forecast, where there are never any hits. A value of 0.5 indicates a random result from the forecast, which can be seen, if the ROC curve sloped diagonally from (0,0) to (1.0,1.0) (see also chapter 14.3.4 in [18]. In practice most forecast systems will produce an AUC between 0.5 and 1.0, with the higher values indicating better forecast performance.

In this example, the AUC values are lowest with a ramp event threshold of 20 MW over one hour with a value of 0.7201 for the HR forecasts and 0.7899 for the LR forecasts. The highest AUC are for the 40MW over three hour threshold with 0.8584 for the LR forecasts and 0.7916 for the HR forecasts. Table 6 indicates that the HR configuration is not providing a more accurate forecast of the occurrence of the events as it scores lower for all event thresholds. This is most likely due to the higher variability in the higher resolution forecasts

Table 6 The "area under the curve" (AUC) summary table for different ramping limits and time windows.

Limit Window	20MW 1h	30MW 3h	40MW 3h	60MW 3h
HR	0.7201	0.7869	0.7916	0.7241
LR	0.7899	0.8351	0.8584	0.8380
$\Delta(HR - LR)$	-0.0043	-0.0053	-0.0049	-0.0028

The range of metrics used in this example provide different perspectives on forecast performance and suggest that one might choose different forecast configurations depending on what forecast attributes are most important for a particular application. The recommended practice (see chapter 15.1.5.1 Evaluation Matrix [9] is to use an evaluation matrix or a cost function (if known) to construct a composite performance assessment to provide a basis for forecast solution selection. For this example, we have an evaluation matrix presented in Table 7. The approach used in this case is to give a score of 1 for better performance on each metric considered in the evaluation and then to multiply this score with an Importance Factor (IF) weight to get the final score. It can be seen that even though the HR ensemble scores worse overall if all metrics have the same importance, it has a higher score if the overall performance, error growth rate and the false alarm rate has higher weight than the other scores. The false alarm rate can be important, if such cases have a high cost in comparison to a correct hit.

Table 7 The result table summarises the scoring for different metrics with an importance factor (IF).

Score	HR	LR	IF weight	HR Final Score	LR Final Score
CRPS	1	0	3	3	0
CRPS leadtime	1	0	4	4	0
BrierScores	0	1	2	0	2
Hit Rate	0	1	1	0	1
False Alarm rate	1	0	2	2	0
Mean Score	0	1	1	0	1
CAL	0	1	1	0	1
DSC	0	1	1	0	1
UNC	-	-	1	-	-
AUC	0	1	1	0	1
SUM	3	6		9	7

4.3 Assessment of instrumentation performance at FINO met mast and Alpha Ventus wind farm

The third use case example focuses on the practical application of the measurement performance assessment recommendations made in section 21.5.1 QC for Wind Forecasting Applications in [21]. This examples illustrates the performance assessment of wind measurements from a measuring platform (FINO1) and wind measurements taken at the offshore wind farm Alpha Ventus.

The Alpha Ventus wind farm is part of the "Research at Alpha Ventus" (RAVE) test field, an initiative supported by the German ministry of economic affairs and climate action [*] since 2006, before the first offshore wind farm was installed. The test field contains two times six wind turbines, a substation and a measurement platform, called FINO1[†]. It is 45 kilometres to the north of Borkum in the German Bight in a water depth of some 30 meters. FINO1 is located in the immediate vicinity of three operating wind farms Alpha Ventus, Borkum Riffgrund I and the westbound TrianelWindpark Borkum and hosts instruments measuring the wind up to 100m.

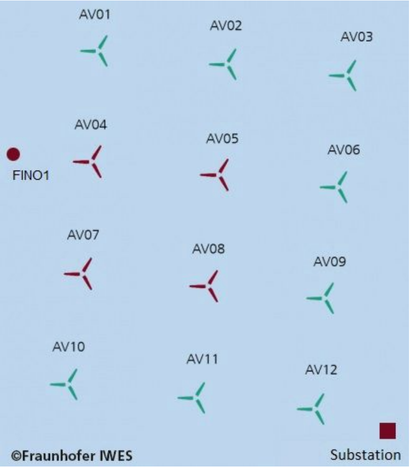

Fig. 9 Turbine arrangements, FINO1 measuring mast and offshore substation at Alpha Ventus wind farm. Graph kindly provided by ©Fraunhofer IWES.

The performance control of wind farms and wind turbines is best conducted in the following four steps:

a) Measuring basic meteorological parameters that can be used to compute power generation output
   - wind speed and direction
   - air temperature
   - barometric pressure
   - relative humidity

b) Conversion of the meteorological parameters into a power output
The best and recommended way is the IEC 61400-12-1 standard on power performance measurements, which is based on a physical formula (Equ. 2, chapter 8 [12])

c) Comparison of power output with measured and forecasted input variables

d) Visual Inspection with ensemble generated percentiles

The first step in this example is the evaluation of the meteorological measurements and generation of statistics that provide information about the fraction of expected data that is not available and whether the agreement between forecast and observation is within an acceptable threshold limit. If the resulting statistical metrics are outside the threshold limits it provides an indication for that there may be issues with the measurements and therefore caution should be taken, when such measurements are used in short-term forecasting.
The first validation considers the physically realistic ranges for the respective measurements. These are listed in Table 8. Table 9 shows the threshold limits for the use case at FINO and Alpha Ventus for the correlation (CORR), bias (BIAS) and mean absolute error (MAE) comparison for four measured variables.

[*] `https://rave-offshore.de/en/about-rave.html`
[†] `https://www.fino1.de/en/`

Table 8 Table with threshold limits for the goodness of data filter of test 1 of the use-case at FINO and Alpha Ventus

Variable	unit	lower limit	upper limit
Wind speed	m/s	0	40
Wind Direction	deg	0	360
Temperature	° C	-40	40
Surface Pressure	PS	800	1.100

Table 9 Variable list and their threshold error limits for the test 2 of use-case at FINO and Alpha Ventus

Variable Number	Variable Name	Min CORR	Max BIAS	Max MAE
1	WindSpeed	0.65	3.0	3.0
2	AirTemp	0.75	2.0	2.5
3	WindDirection	0.55	13.0	20.0
4	AirPressure	0.90	50.0	85.0

Table 10 shows an example of results from the evaluation period for 6 turbines (AV07 .. AV12) and the substation (SUB) that aggregates the power from the individual turbines and connects it with the transmission lines that bring it onshore. The wind speed for SUB is taken from the FINO1 mast at 100m. Test 1 encompasses wind speed, temperature, wind direction, surface pressure with the binary indicators 1 for accepted quality and 0 for failing the requirements set out in Table 9.

Table 10 Variable list and their threshold error limits for the use-case at FINO and Alpha Ventus

Site	Test 1		Test 2	(Tab9)		Description
ID	(Tab8)	WS	T2m	WDIR	PS	
AV07	1111	111	111	111	111	all tests ok
AV08	1111	111	111	111	111	all tests ok
SUB	1110	111	111	111	000	PS fails all tests
AV09	1101	111	111	100	111	WD fails except for WD(BIAS) OK
AV10	1101	111	111	101	111	WD fails except for WD(MAE) OK
AV11	1010	111	000	111	110	T fails on all
AV12	1001	111	000	101	111	T fails and WD(MAE) fails

The analysis period for the first evaluation set was from June to October 2021. It should be noted that for a robust statement about an instrumentation or in real-time environments, where the measurements are used for real-time forecasting purposes, a minimum period of one year is ideal (for forecast training and calibration), but often not available. An alternative – especially in real-time applications – is to regularly test the performance, e.g. on a quarterly basis (see e.g. [22]).

Tables 11-13 show another evaluation set for different periods for individual turbines AV07, AV08 and the SUB for the first three quarters of 2021. In this evaluation set, two more requirements were added in order to identify performance, but also availability of the data in real-time: (Test 3): the power output error derived with a standard power curve computation according to IEC 61400-12-1 [23] needs to be 5% lower with measured wind speed than with forecasted wind speed and **Test 4** the delivery rate has to be $> 98\%$.

Table 11 Category "Bad Data || Missing data + Requirement 2: Improvement < 5%" for 1^{st} quarter of 2021. Test 3 evaluates the improvement of produced power when computed with measured instead of forecasted wind. Test 4 verifies the availability of the measured data.

Site	Test 1		Test 2			Test 3	Test 4
		WS	T2m	WDIR	PS	>5%	
AV07	0101	0111	1111	1001	1111	0	47.7

Table 12 Category "Bad DATA and MiSSING DATA" for 2^{nd} quarter 2021. Test 3 and 4 as in Table 11

Site	Test 1		Test 2			Test 3	Test 4
		WS	T2m	WDIR	PS	>5%	
AV08	1001	1111	0001	0001	1111	6.57	10.6
AV07	1001	1111	0001	0001	1111	6.14	11.4

Table 13 Category "Good data" for 3^{rd} quarter of 2021. Test 3 and 4 as in Table 11

Site	Test 1		Test 2			Test 3	Test 4
		WS	T2m	WDIR	PS	>5%	
SUB	1111	1111	1111	1111	1111	2.19	99.8

To summarise the use case of verifying and quality assessing meteorological measurements at a wind farm, we can conclude that the recommendations made in the the IEA Wind Recommended Practice guideline part four [24] regarding quality assessment of measurements are important considerations, when using measurements for forecast assimilation (adaptation) in a real-time environment, for situational awareness or grid related down-regulation. Unless the data is reliable, the damage done to a forecast can be greater than the potential improvement.

In our example evaluation period, the tests revealed the following about the measured data:

- observations correlate well with model level 3 at around 100m
- the observations at heights 90m and 100m show 18% missing data
- the best data coverage is between 10m and 40m
- if signals stall, they stall over most levels, some with phase shifts of 2-3 time intervals
- missing data signals are often associated with precipitation and/or high wind speeds

The two latter observations are consistent with literature (e.g. [25–27]). As forecaster, as well as operator of instrumentation and end-user of the data, it is crucial to understand the potential issues that can arise from a lack of data, wrong or misleading data. Such analyses can hence provide valuable guidance to both those that operate the instrumentation, the forecaster and end-user and usually have well balanced cost-benefit ratio.

5 Summary

The Recommended Practice (IEA-RP) developed by the IEA Wind TCP under Tasks 36 and 51 provides a comprehensive set of guidance for the selection and implementation of optimal forecast solutions for specific applications. This paper provides a brief overview of the contents of the four parts of the IEA-RP. It also introduces the development and availability of a companion forecast evaluation tool, called WE-verify-prob [10], that is a resource for the evaluation of forecast performance with a focus on the assessment of probabilistic forecasts. However, the focus of the paper is a presentation of three use case examples of how specific guidelines of the IEA-RP can be applied.

The three use case examples are (1) evaluation of probabilistic 10-m wind forecasts for a standard meteorological measurement site in Denmark and (3) analysis of the performance of probabilistic forecasts of the wind power generation fed into a substation in Northwest Ireland and(3) assessment of the quality of wind measurement data at an off-shore meteorological measurement tower in the German Bight. The examples illustrate the importance of using a set of evaluation metrics rather than a single metric to assess performance. They also demonstrate the potential value of constructing a metric matrix to construct a summary metric that includes a weighting of the relative importance of different forecast performance attributes for a specific application. That is, different applications should have different attribute (i.e. metric) weights in order to obtain a composite score that is most relevant for a specific application. In the example presented in this paper, a forecast configuration that had lower performance in the majority of the metrics was selected, because it had higher performance for the metrics that were deemed to be most important for the application.

6 Acknowledgements

The authors would like to thank Jethro Browell for his support in the WE-verify-prob software development. Corinna Möhrlen, Mathias Blicher Bjerregård and Gregor Giebel are supported by the Danish EUDP project IEA Wind Task 51 Danish Consortium under contract no. 134-22015.

References

[1] A. F. A. J. S. M. M. Hodge, B-M and D. Mcreavy, “The value of improved short-term wind power forecasting,” National Renewable Energy Laboratory, Tech. Rep., February 2015. [Online]. Available: \url{https://www.nrel.gov/docs/fy15osti/63175.pdf}

[2] K. Orwig, B.-M. Hodge, G. Brinkman, E. Ela, M. Milligan, V. Banunarayanan, S. Nasir, and J. Freedman, “Economic evaluation of short-term wind power forecasts in ercot: Preliminary results; preprint,” 9 2012. [Online]. Available: https://www.osti.gov/biblio/1051163

[3] J. C. H. F. W. S. B.-M. H. C. D. P. P. J. M. G. K. C. M. Giebel, G., “Iea wind task 36 wind energy forecasting,” 2018. [Online]. Available: https://www.iea-wind.org/task36/publications

[4] C. Möhrlen, J. W. Zack, and G. Giebel, *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, 1st ed., ser. Wind Energy Engineering. Academic Press, 2022. [Online]. Available: https://www.sciencedirect.com/book/9780443186813/

[5] ——, “Introduction,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, p. 1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000106

[6] ——, “Introduction,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, p. 77. [Online]. Available: \url{https://www.sciencedirect.com/science/article/pii/B9780443186813000167}

[7] ——, “Introduction,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, p. 105. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000222

[8] ——, “Introduction,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 185–186. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000283

[9] ——, “Chapter fifteen - best practice recommendations for forecast evaluation,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 147–184. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000271

[10] “Appendix g - validation and verification code examples,” in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp.

321–322. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000428

[11] C. Möhrlen, J. W. Zack, and G. Giebel, "Chapter ten - best practice recommendations for benchmarks/trials," in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 101–103. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000210

[12] ——, "Chapter eight - conducting a benchmark or trial," in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 89–96. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000192

[13] T. Dimitriadis, T. Gneiting, and A. I. Jordan, "Stable reliability diagrams for probabilistic classifiers," *Proceedings of the National Academy of Sciences*, vol. 118, no. 8, p. e2016191118, 2021. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2016191118

[14] J. E. Matheson and R. L. Winkler, "Scoring rules for continuous probability distributions," *Management science*, vol. 22, no. 10, pp. 1087–1096, 1976.

[15] F. Krüger, S. Lerch, T. Thorarinsdottir, and T. Gneiting, "Predictive inference based on markov chain monte carlo output," *International Statistical Review*, vol. 89, no. 2, pp. 274–301, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12405

[16] M. Zamo and P. Naveau, "Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts," *Math Geosci*, vol. 50, pp. 209–234, 2018.

[17] c. Timo Dimitriadis [aut], Alexander I. Jordan [aut, *reliabilitydiag: Reliability Diagrams Using Isotonic Regression*, 2022, r package version >. [Online]. Available: https://CRAN.R-project.org/package=reliabilitydiag

[18] C. Möhrlen, J. W. Zack, and G. Giebel, "Chapter fourteen - assessment of forecast performance," in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 125–145. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978044318681300026X

[19] Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Elsevier B.V., 2011.

[20] A. H. Murphy, "Skill scores based on the mean square error and their relationships to the correlation coefficient," *Monthly Weather Review*, vol. 116, no. 12, pp. 2417 – 2424, 1988. [Online]. Available: https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml

[21] C. Möhrlen, J. W. Zack, and G. Giebel, "Chapter twenty-one - assessment of instrumentation performance," in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 251–275. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000349

[22] C. Möhrlen, D. Ó Foghlú, S. Power, G. Nolan, K. Conway, E. Lambert, and J. Ging, "Eirgrid's met mast and alternatives study," *IET Renewable Power Generation*, vol. 16, no. 9, pp. 1941–1954, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/rpg2.12502

[23] I. E. C. (IEC), "IEC standard 61400-12-1:2017 power performance measurements of electricity producing wind turbines," International Electrotechnical Commission, Tech. Rep., 2017.

[24] C. Möhrlen, J. W. Zack, and G. Giebel, "Chapter twenty-two - best practice recommendations," in *IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions*, ser. Wind Energy Engineering, C. Möhrlen, J. W. Zack, and G. Giebel, Eds. Academic Press, 2023, pp. 277–299. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443186813000350

[25] S. Drechsel, G. J. Mayr, J. W. Messner, and R. Stauffer, "Wind speeds at heights crucial for wind energy: Measurements and verification of forecasts," *Journal of Applied Meteorology and Climatology*, vol. 51, no. 9, pp. 1602–1617, 2012. [Online]. Available: \url{https://journals.ametsoc.org/view/journals/apme/51/9/jamc-d-11-0247.1.xml}

[26] C. Draxl, L. K. Berg, and L. B. et. al, "The verification and validation strategy within the second wind forecast improvement project (wfip 2)," National Renewable Energy Laboratory, Tech. Rep., 2019. [Online]. Available: https://www.nrel.gov/docs/fy20osti/72553.pdf

[27] E. o. m. d. Joint Committee for Guides in Metrology, *The role of measurement uncertainty in conformity assessment*, 2012.