RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

- Part 2: DESIGNING AND EXECUTING FORECASTING BENCHMARKS AND TRIALS -

2. EDITION

Accepted by the Executive Committee of the International Energy Agency Implementing Agreement in January 2022

Prepared by IEA Wind Task 36 and 51

Document Version: 2.0
February 2022

Contents

Preface

This 2^nd edition of this recommended practice document is the result of a collaborative work that has been edited by the undersigning authors in alignment with many discussions at project meetings, workshops and personal communication with colleagues, stakeholders and other interested persons throughout the phase 2 of the IEA Wind Task 36 (2018-2021) as part of workpackage 2.1 and 3.1.

The editors want to thank the Danish EUDP for funding the work under the Grant Number 64018-0515, and everybody that contributed in our meetings, workshops and sessions to the discussions, provided feedback or other input throughout the past 6 years.

IEA Wind Task 36, December 2021

Editors and Authors:
Dr. Corinna Möhrlen (WEPROG, Denmark) <com@weprog.com>
Dr. John Zack (UL AWS Truepower, USA) <john.zack@ul.com>

Contributing Authors:
Dr. Craig Collier (Energy Forecasting Solutions, USA)
Dr. Aidan Tuohy, EPRI, USA
Dr. Jakob W. Messner (MeteoServe Wetterdienst, Austria)
Dr. Jeffrey Lerner (ENFOR, Denmark)
Dr. Jethro Browell (University of Glasgow, United Kingdom)
Dr. Justin Sharp (Sharply Focused, USA)
Mikkel Westenholz (ENFOR, Denmark)

Supported by:
Operating Agent Dr. Gregor Giebel (Danish Technical Univerity, DTU Wind, DK)

Chapter 1

Background and Objectives

1.1 BEFORE YOU START READING

This is the **second part** of a series of three recommended practice documents that address the selection, development and operation of forecasting solutions. It addresses benchmarks and trials in order to test or evaluate different forecasting solutions against each other and the fit-for-purpose.
The **first part** *Forecast Solution Selection Process* addresses the selection and background information necessary to collect and evaluate when developing or renewing a forecasting solution for the power market. The **third part**, *Forecast Solution Evaluation*, which is the current document, provides information and guidelines regarding effective evaluation of forecasts, forecast solutions and benchmarks and trials. The **fourth part**, *Meteorological and Power Data Requirements for real-time forecasting Applications*, provides guidance for the selection, deployment and maintenance of meteorological sensors and the quality control of the data produced by those sensors with the objective of maximising the value of the sensor data for real-time wind and solar power production forecasting.
If your main interest is (1) selecting a forecasting solution, (3) verifying the quality of your forecast solution, or (4) setting up meteorological sensors or power measurements for real-time wind or solar power forecasting, please move on to part 1, 3 or 4 of this recommended practice guideline to obtain recommendations on any of these specific issues, respectively.

It is also recommended using the table of contents actively to find the most relevant topics.

1.2 BACKGROUND

The effectiveness of forecasts in reducing the variability management costs of power generation from wind and solar plants is dependent upon both the accuracy of the forecasts and the ability to effectively use the forecast information in the user's decision-making process. Therefore, there is considerable motivation for stakeholders to try to obtain the most effective forecast information as input to their respective decision tools.

This document is intended to provide guidance to stakeholders on a primary mechanism that has been used extensively in the past years to assess the accuracy of potential forecasting solutions: benchmarks and trials.

This guideline focuses on the key elements to carry out a successful trial or benchmark and on typical pitfalls. It will also provide recommendations as to when it is beneficial or too risky or expensive in terms of resources to carry out a trial or benchmark.

1.3 DEFINITIONS

The two main terms and concepts "trial and benchmark" that are used in this recommended practice shall be defined in the following. Note, the focus has been on forecasting processes in the power industry and the definition may not have a completely general character to be applied to other areas of business. Additionally, it should be noted that "forecasting trials and benchmarks" will be abbreviated with "t/b" throughout this document for simplicity.

1.3.1 Renewable Energy Forecast Benchmark

A "renewable energy forecast benchmark" is in this document defined as an exercise conducted to determine the features and quality of a renewable energy forecast such as wind or solar power. The exercise is normally conducted by an institution or their agent and multiple participants, including private industry forecast providers or applied research academics.

1.3.2 Renewable Energy Forecast Trial

A "renewable energy forecast trial" is in this document defined as an exercise conducted to test the features and quality of a renewable energy forecast, such as wind or solar power. This may include one or more participants and is normally conducted by a private company for commercial purposes. A trial is a subset of a Renewable Energy Forecast Benchmark.

1.4 Objectives

The guidelines and best practices recommendations are based on years of industry experience and intended to achieve maximum benefit and efficiency for all parties involved in such benchmark or trial exercises. The entity conducting a trial or benchmark taking the recommendations provided in this guideline into consideration will have the following benefits:

1. Being able to evaluate, which of a set of forecast solutions and forecast service providers (FSP) fits best the need, specific situation and operational setup
2. Short term internal cost savings, by running an efficient t/b
3. Long term cost savings of forecast services, by following the trial standards and thereby help reduce the costs for all involved parties

In the discussion of the process of obtaining the best possible forecasting solution, there are a number of terms and concepts that are used. Several of the key terms and concepts are defined in the following.

Note, these definitions are kept as general as possible with a focus on forecasting processes in the power industry and may not have such a completely general character to be applied to other areas of business.

Chapter 2

Initial Considerations

Key Points
This section is targeted to the task of engaging a forecast service provider (FSP) and how to navigate through the vast amount of information.

2.1 Deciding whether to CONDUCT a Trial or Benchmark

The most important initial consideration when planning a forecasting trial or benchmark (t/b) is to be clear about the desired outcome.
The following tables provide information about the benefits and drawbacks of conducting a t/b as a key part of the selection process. Before a decision is made to conduct a t/b, it is recommended to go through these tables and determine, if the effort is warranted.
A possibly attractive alternative approach for a forecast user that wishes to evaluate a set of forecast solutions for their ability to meet the user's needs is to engage an independent trial administrator. An experienced and knowledgeable administrator can act as a neutral third party and advocate for both the vendors and the end-users in the design and execution of a t/b and the evaluation and interpretation of the results. Such an arrangement builds trust in the process among all parties.
An effective administrator can take the requirements from the user and ensure they are realistically incorporated into the trial design. There obviously is a cost to engage such an administrator, but it may be more cost-effective for the user and generate more reliable information for the user's decision-making process.

2.2 Benefits of Trials and Benchmarks

Table 2.1: Decision support table for situations in which trials/benchmarks are determined to be beneficial

Situation	**Benefit**
Real-time trial for an entire portfolio	High cost but information gain is greater and more representative; provides the best estimate of the error level and which solution/FSP is best for the target applications
Real-time trial for a selected number of sites	Lower cost but still a substantial information gain if sites are well selected; provides a reasonable idea about the error level and a good indication of which solution/FSP fits is best for the target applications
Retrospective benchmark with historic data for a specific time period separate from a supplied training data set	Low cost: In multi-FSP systems, the error level of an additional FSP is secondary, while the correlation with other FSPs determines whether the additional FSP improves the overall error of a multi-FSP composite forecast
Blind forecast without historic measurements	Test to get an indication of the accuracy of forecasts from an FSP in the upstart phase of a project, where no historical data are available. Excludes statistical methods, which need historical data. An inexpensive way to get an indication of forecast accuracy for larger portfolios (> 500MW), where measurement data handling is complex. NOTE: There is an inherent risk that the result may be random and FSP use different methods for blind forecasting and forecasting with measurement data. See also Table 2.2 for limitations of this approach.

2.3 Limitations with Trials and Benchmarks

Table 2.2: Decision support table for situations in which trials/benchmarks are determined to contain limitations and a t/b is not recommended.

Situation	**Limitation**	**Recommendation**
Finding best service provider for large portfolio (> 1000MW) distributed over a large area	Trial for entire portfolio is expensive for client and FSP in terms of time and resources.	Limiting scope of trial limits representativeness of results for entire portfolio. RFI and RFP in which FSP's methods are evaluated and the use of an incentive scheme in the contract terms provides more security of performance than a limited trial.
Finding best service provider for a medium sized portfolio (500MW< X < 1000MW) over a limited area	Trial for entire portfolio is expensive for client and service provider in terms of time and resources.	Limiting scope of trial limits representativeness of results for entire portfolio. RFP in which FSP's methods are evaluated. Design of a system that enables an easy change of FSP and use if an incentive scheme is more a more cost effective approach than a trial.
Finding best service provider for small sized portfolio (< 500MW)	Trial for entire portfolio usually requires significant staff resources for about 6 months	Trial is feasible, but expensive. Difficult to achieve significance on target variable in comparison to required costs and expenses – trial costs makes solution more expensive. Less expensive to setup an incentive scheme and a system where the FSPs can be changed relatively easily.
Finding best service provider for micro portfolio (< 100MW) or single plants	Cost of a trial with many parties can easily be higher than the cost of a 1-year forecasting contract.	Time for a trial can delay operational forecast utilization by up to 1 year! Select FSP based on an evaluation of methods and experience.

Situation	Limitation	Recommendation
Design a system that enables an easy change of FSP and use an incentive scheme for FSP performance	Power marketing	Best score difficult to define, as sale of energy is also dependent on market conditions and a statistical forecast performance score such as RMSE or MAE does not reflect the best marketing strategy More efficient and timely to perform back test of historical forecasts combined with historical prices, or make a strategic choice with an performance incentive.
Market share of FSP in a specific power market is high	FSP monopolies in a specific power market mean that forecast errors are correlated and hence increase balancing costs.	Ask about the market share of a provider and do not choose one with a share > 30% as the only provider!
Blind forecasting, i.e. no historic measurement data available	Without measurements the value of a trial is very limited due to the significant improvement from statistically training forecasts and the importance of recent data for intra-day forecasts	Evaluation can only be meaningfully done for day- ahead or longer forecasts.
Some FSP may us different methods for forecasting with and without historic data (statistical methods need historical data to function!)	Results are limited to testing quality on upstart phase of new projects, where no historical data exist (see also Table 1).	For single sites, the benefits of training are so large (>50% of error reduction at times) that blind forecasting is not recommended. For larger portfolios it can provide an indication of quality - for physical conversion methods only!

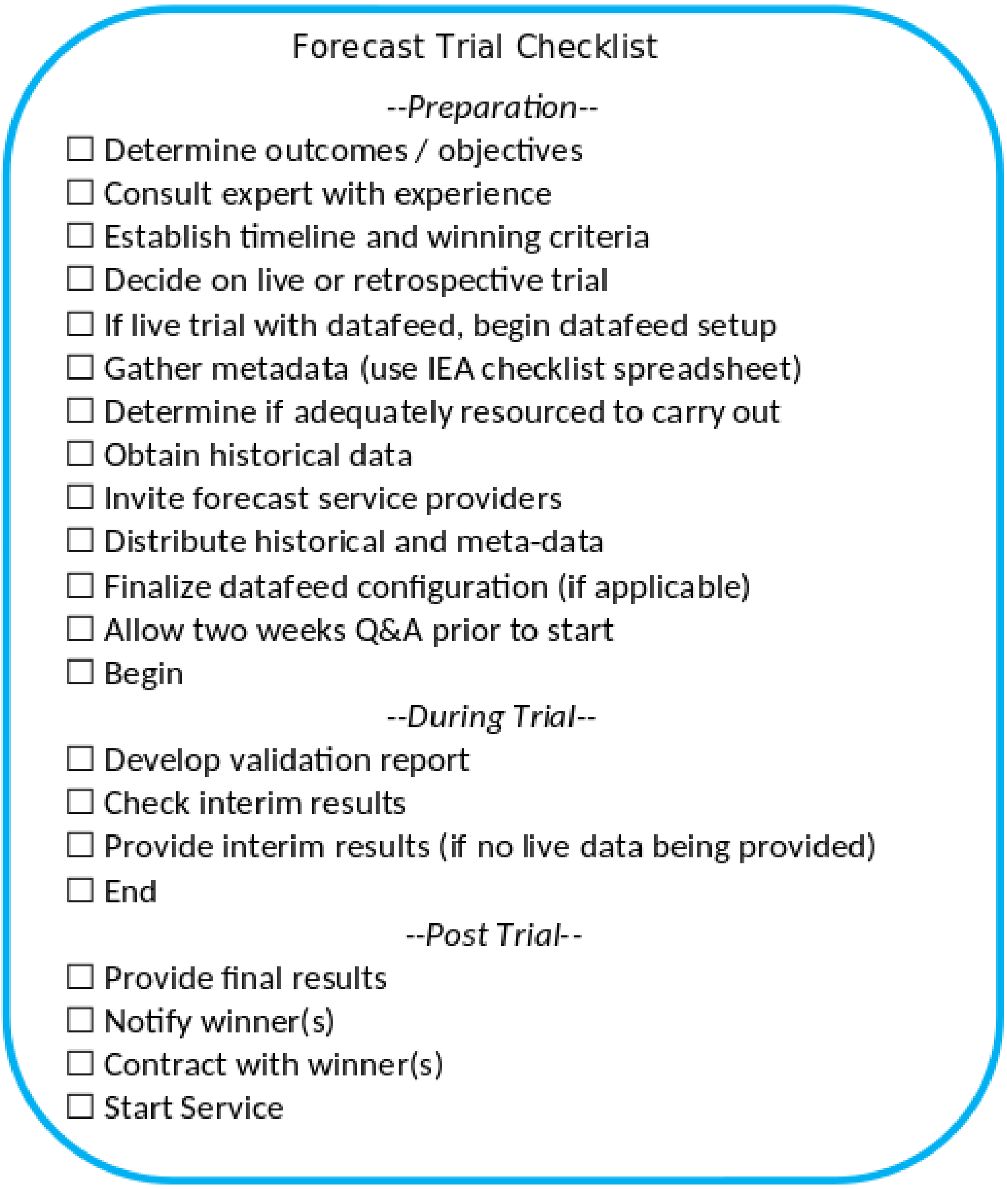
2.4 Time lines and Forecast periods in a Trial or Benchmark

Time lines and forecast periods need to be set strictly in a trial or benchmark in order to achieve a fair, transparent and representative exercise.
The following time lines should be considered:

1. Start and stop dates of the t/b must be fixed
2. Start and stop dates must be the same for all FSPs
3. Pre-trial setup and test dates for IT infrastructure (including any required security protocols) for trial must be specified and enforced
4. Delivery times of forecasts must be set and enforced
5. Forecasts for periods with missing forecasts from one FSP must be excluded for all FSPs

2.5 1-PAGE "Cheat sheet" Checklist

The following checklist is provided to help trial organizers save time, apply best practices, and avoid common pitfalls when designing and executing forecast trials. It has been compiled by leading forecast vendors and researchers with many years experience.



Forecast Trial Checklist

--Preparation--

- ☐ Determine outcomes / objectives
- ☐ Consult expert with experience
- ☐ Establish timeline and winning criteria
- ☐ Decide on live or retrospective trial
- ☐ If live trial with datafeed, begin datafeed setup
- ☐ Gather metadata (use IEA checklist spreadsheet)
- ☐ Determine if adequately resourced to carry out
- ☐ Obtain historical data
- ☐ Invite forecast service providers
- ☐ Distribute historical and meta-data
- ☐ Finalize datafeed configuration (if applicable)
- ☐ Allow two weeks Q&A prior to start
- ☐ Begin

--During Trial--

- ☐ Develop validation report
- ☐ Check interim results
- ☐ Provide interim results (if no live data being provided)
- ☐ End

--Post Trial--

- ☐ Provide final results
- ☐ Notify winner(s)
- ☐ Contract with winner(s)
- ☐ Start Service

Figure 2.1: "Cheat sheet" Checklist

Chapter 3

Conducting a Benchmark or Trial

Key Points:
Deterministic trials have become an established way to test different forecast venders or test the compatibility and benefits of combining various forecast methods in the forecast solution selection process. Such trials are complicated and the required ressounces to conduct fair, transparent and representative results are often uncerestimated. In order to generate valuable results, such trials need to follow a specific structure, which is characterised by three phases:

- *Phase 1: Preparation*
- *Phase 2: During Trial*
- *Phase 3: Post Trial*

These three main phases of a trial exercise, preparation ahead of the trial, actions during the trial, and post-trial follow up are described in detail in the following.

3.1 Phase 1: PREPARATION

The time required for the pre-trial preparation is significant and should not be underestimated to insure a successful outcome. If the operator of the trial has no experience in renewable energy forecasting or running a t/b, it would be prudent to contact an experienced individual, organization or forecast provider to obtain feedback on what can reasonably be accomplished given the target time line and objectives. Part 1 of this recommended practice contains a decision support path that may be useful for determining the proper course of action.

3.1.1 Key Considerations in the Preparation Phase

Once the objectives of the t/b are known (see Section 1.1 Background and 1.2 Objectives), there are some key decisions to be made that will play a major role in determining the complexity of the trial. They are:

1. **Choice of forecast horizon:**
   Are forecast horizons less than 6 hours operationally important? If the answer is "no", establishing a live data feed may not be necessary. Although there are advantages of running a trial with a live data feed, it is one of the most time consuming aspects of trial preparation. Are forecast lead times greater than “day-ahead” operationally important? If the answer is no, this will reduce the volumes of data that need to be processed, saving time and resources. If many lead times are of operational importance, consider that the performance of different providers will likely vary across lead times, therefore, different lead times, e.g. hour-ahead, day-ahead and week-ahead, should be evaluated separately.

2. **Weather conditions for the exercise:**
   Will the benchmark take place during periods of more difficult to predict weather conditions that reflect the organization’s difficulties in handling renewable generation, e.g. windy or cloudy periods? The answer here should be "Yes" to insure the sample size of harder-to-forecast events is sufficient. If the answer is "No", the trial operator should strongly consider doing a retrospective forecast (also known as "backcast") that includes the types of conditions that are critical for the user’s application.

3. **Historical data/observations for the exercise:**
   For locations in which there are significant seasonal differences in weather conditions and the associated renewable generation levels and variability, it is best to provide 12 months or more of historical data from the target generation facilities to the FSPs for the purpose of training their forecast models. However, if it is not feasible to make this amount of data available or if the target location does not exhibit much seasonal variation, most FSPs can typically train their forecast models reasonably well with 3-6 months of on-site historical observations.

   It should be noted that advanced machine learning methods often exhibit significantly greater performance improvement over less sophisticated methods as the training sample size increases. Thus, FSPs that employ the latest and most advanced machine learning prediction tools may not be able to demonstrate the ultimate value of their approaches, if only short historical data sets are provided. If 6-12 months of data are not available, the trial operator might consider another location or conduct a longer trial on the order of 4-6 months to monitor forecast improvements over time as more data becomes available to the FSPs to improve the quality of the training of their prediction models.

In general it is recommended that the t/b operator should provide a dataset of the typical length that is available data for the application that is the target of the t/b. If more historical data is available for a t/b than in the typical application, care should be taken in the evaluation of methods, as e.g. machine learning methods might outperform e.g. physical methods in the trial, but perform worse in the real application due to the benefits associated with the longer data sets.

4. **Representativeness:**
   Is the benchmark location representative from a wind-climatology perspective of the scope of locations for which the operator will ultimately require operational forecast services? That is, the trial operator should select a location that is needed for subsequent forecasting or a location with a similar climatology. It should also be noted, that if different vendors provide forecasts for only one single site, that forecast performance has a certain random quality over shorter periods of weeks or months due to the non-linear behaviour of weather conditions and associated variable performance of NWP and power conversion models (see e.g. Collier[5] for an example demonstrating this challenge). Additionally, forecast performance exhibits a significant "aggregation effect". That is, the magnitude and patterns of forecast errors vary substantially depending on the size and composition of the forecast target entity. Thus, the characteristics of forecast errors for an individual turbine, a single wind park and a portfolio of wind parks will typically be quite different, and the forecast evaluator should be very careful when inferring forecast performance characteristics from one scale of aggregation (e.g. a single wind park) to a different scale (e.g. a geographically diverse portfolio of wind parks) (see also part 3 of this recommended practice for more details on evaluation methods).

5. **Metrics:**
   Are the metrics that will be used to evaluate the forecasts meaningful to the success of my project? There are a wide variety of well-documented error metrics that penalize forecast errors differently. For example, root mean squared error penalizes large errors more than small errors. It is important to choose a metric, or set of metrics, that reflects the value of an improved forecast to the user's application and can discriminate between different forecast solutions. Please refer to part 3 of this recommended practice for details on metric selection.

3.1.2 Metadata Gathering in the Preparation Phase

Details of the forecast trial, such as location and capacity of the target generator, are required by all FSPs and comprise the trial Metadata. Appendix A "Metadata Checklist" provides the information that is typically needed by FSPs for participation in a trial and is designed to be used as a spreadsheet form that is completed during the preparation phase of a t/b. This should also include the desired format (filename and content) of the forecasts you'll be comparing. The best way to communicate the forecast file format to multiple FSPs is to provide an example file.

3.1.3 Historical Data Gathering in the Preparation Phase

On-site observations of power production or the renewable resource (e.g., irradiance or wind speed at hub height) are critical for helping the FSPs statistically "train" their forecast models and thus reduce error and bias in the forecasts. Good quality data is critical. "Good quality" means that the data does not, for example, contain many gaps or unrepresentative values. Curtailed power data should be accompanied by plant availability or a curtailment flag.

Data time intervals should be regular and there should be a clear documentation of the units, how the observations were averaged, the time zone of the data, and whether there's a shift in time due to daylight savings time. Appendix A of this document has a concise list of the necessary historical data attributes required to efficiently start a t/b.

3.1.4 IT/Data Considerations in the Preparation Phase

Most organisations have constraints on the amount of IT resources available for a t/b. Therefore, it is best to plan ahead or keep the sending and receiving of data very simple. The primary IT issue is typically the selection and setup of data formats and communication protocols that will be used for the t/b operator to send data to the FSPs and for the FSPs to send forecasts to a platform designated by the t/b operator.

Data formats:

There are many possibilities for data formats, which range from a simple text file with comma separated variables (CSV) to more sophisticated XML or openAPI formats. Similarly, there are a wide range of communication protocols that can be used. These range from the relatively simple Secure Shell File Transfer Protocol (SFTP) to more sophisticated web service or API structures. The more sophisticated structures have advantages and there are many IT companies and resources that support these structures but they almost unavoidably increase the complexity of the setup.

Unless adequate IT resources or knowledge are available for all participants (especially the operator) it is recommended that simple data formats and communication resources be employed for a t/b. This typically means the use of the CSV data format and an SFTP data communications protocol.

Live trial considerations:
If a live trial is planned (most common), but real-time data will not be made available to the FSPs, then a place for each FSP to send forecast files will need to be setup. One of the metrics that is often used to evaluate an FSP is the timeliness of forecast delivery. In this case, it is important that a mechanism to verify the time of delivery be established. If real-time data is provided by the t/b conductor, it is typically easiest to create a common password-protected file server directory from which FSPs can download the data via a protocol such as SFTP. Another approach is to use SFTP to push data files to each FSP. This typically requires more effort, especially for the t/b operator.
Historical data can be provided to FSPs in the same data format via the same communication protocol. However, it often requires a SCADA engineer or expert on third party software to extract the historical data for the SCADA (or other) data archive.

Legal Agreements:
Another often-overlooked data-related issue is the legal agreements required to disseminate data from possibly multiple data provider entities (e.g. the wind facility owners/operators) to multiple data user entities (e.g. the FSPs in the t/b). This may be relatively simple in cases in which the user (such as a generator fleet operator) owns all the data and is willing to make it available for the t/b with few restrictions. However, it be a very complex and time consuming process in cases in which the user (e.g. a system operator) does not own the data and merely serves as a conduit from the multiple data owners with different data dissemination restrictions to the data users.
In such cases, the process of formulating and executing the required legal documents (such as non-disclosure agreements (NDAs)) can cause substantial delays in the initiation of a t/b and perhaps even change its scope.
See Appendix B for example formats in csv and xml.

3.1.5 Communication in the Preparation Phase

Transparency:
Anonymising the FSPs for all communication is considered a best practice as it ensures transparency of the available information, promotes competition and entry from smaller FSPs trying to become more established in the industry. Commu-

nication via email therefore should always be consistent with blind copies to all FSPs.
Consistency:
Consistent in this context means always sending and sharing emails with the same group of FSP users. Common information sharing engenders trust and the perception of fairness in the benchmark or trial process. In the preparation phase, it is not uncommon that the FSPs will have questions that could affect how the trial is conducted.
For this reason, it is recommended to have a 2-week question and answer period before the official start date to allow FSP participants to ask questions that then can be answered in a living document that contains all questions and answers up to the present time. All participants should be notified whenever this document is updated.
Frequency:
The importance of frequent and clear communication cannot be overstated when conducting a t/b. Not only will the t/b operator receive the most accurate forecasts, it will make it much easier the next time a t/b is executed to gage the state-of-the-art in forecasting technologies and features.

3.1.6 Test run in the Preparation Phase

It is recommended to that a minimum of one-week is allocated for a test period before the official start date of the t/b to identify and remove any technical issues that could invalidate forecast results. This helps to improve the likelihood that all results can be included in the final validation calculations without the need for omitting the first part of the t/b.

3.2 Phase 2: DURING BENCHMARK/TRIAL

Verification & Validation Report preparation Often the most successful forecast provider is one that can show steady improvement over time. Providing an interim validation report will not only prepare the trial operator for the final validation report but will give important feedback to the FSPs – not only throughout the trial or benchmark, but also in the daily operations.
Validation Strategy:
Part 3 of this recommended practice provides information about validation and verification that incentivices the FSP, where it is beneficial for the end-user.
Verification strategy:
In Draxl 5, a verification and validation strategy is described that emphasizes that verification of validation code is an essential part of a validation. In the case of a trial or benchmark, it is recommended that the verification strategy and the input

data for the validation is shared with the FSP. In that way, the verification code is tested as recommended by Draxl 5 and there is transparency on the results. If the FSPs result differs from the end-user's result, the errors can be detected and solved.

3.2.1 Communication during the T/B

In a well-designed t/b, most of the communication between the trial operator and FSPs should be during the pre-trial period. However, issues often arise especially during a live trial with a real-time data feed. It may be helpful to all t/b participants to establish an open forum during the first part of the live t/b period (e.g. the first 2 weeks) to provide a way to effectively and uniformly resolve all issues early in the t/b period However, it is strongly recommended that if any attributes of the t/b are changed at any point during the live part of the t/b, the changes should be communicated to all participants immediately as they might require action on the FSP's part.

Examples might include: changing the forecast validation metric, if there are unreported outages that should be omitted for future model trainings, or if the location of the data feed or forecast file destination has changed. It should be emphasized that all communications related to the t/b should be distributed to all FSPs without exception. Additional communication with individual FSPs (including forecast incumbents) can be interpreted as bias on the part of the operator of the t/b and in some cases may actually bias the t/b result due to information that impacts forecast design, production or delivery not being equally available to all FSPs.

3.2.2 Forecast Validation and Reporting during the T/B

Forecast validation reports are often compiled during the t/b. With forecast data coming in at regular intervals, the t/b operator has real data to feed into the validation report. If the t/b has a duration of several months (i.e., >3 months), it is recommended to provide at least one interim report to FSPs that include anonymized results from all FSPs. This benefits the trial operator as errors in the evaluation process or the report generation can be flagged earlier and ways to make the report generation more efficient can be realized. The interim report benefits the FSPs as course-corrections can be made during the t/b to improve the forecasts.

If there are several FSPs participating, efficiencies can be realized by automating part or most of the validation metrics especially as the forecast file format should be the same from all FSPs.

3.3 Phase 3: POST TRIAL OR BENCHMARK

The post trial phase is an important aspect of the t/b because FSP selection will likely occur during this phase based on the criteria set out at the start of the t/b. (see recommended practices part 1 on "evaluation of services and decision support").

3.3.1 Communication at the end of the T/B

If the trial operator hasn't already done so, an email should be sent within a week before the end date of the t/b to alert FSPs that the end of the trial is near and to communicate the timeline for sharing results and re-iterate the specifications of the FSP selection process.

3.3.2 Forecast Validation and Reporting at the end of the T/B

If an interim report was provided during the trial, then the final report can either be an updated version of the validation report expressing the bulk metrics or appended month-by-month forecast validation results. For transparency and to promote further forecast improvements, it is recommended that the t/b operator share the anonymized forecast results from each FSP at the time-interval frequency that forecasts were being made at (e.g., hourly). This will help FSPs discover where forecasts are similar or different from the competition which may spawn improved methodologies.

Chapter 4

Considerations for Probabilistic Benchmarks and Trials

Key Points:
Testing, verification and validation of probabilistic forecast methods and forecast solutions need to be handled fundamentally different than deterministic methods. The latter can be aggregated, combined and compared and has in the past been mostly used to foster improvements on basic statistic metrics.

Probabilistic forecast methods theoretically can be used as deterministic forecasts as well, for example a mean, a percentile or quantile forecast with a specific target. In that case, the deterministic evaluation can be used as described in sections 3.1, 3.2 and 3.3.

When dealing with uncertainties in the forecast process chain, probabilistic forecasts have value beyond a deterministic forecast and hence, the evaluation is different (see 4.1). Such forecasts serve a different purpose and can be compared, but not - in a straight forward or easy way – aggregated or combined in the same way to improve average forecast metrics.

In trials and benchmarks with probabilistic solutions, the verification should be done by method (see examples in Table **??**) and in most cases with event based verification metrics such as:

(i) "Event evaluation"
Examples are categorial event analysis with contingency tables, critical success index (CSI), measuring the ratio of correct event forecasts to the total number of forecasted and observed events (see section 4.2)

(ii) "Cost or Loss Functions"
Such functions measure the sensitivity of a user's application to the forecast error which for example can be wether the observations is within the forecasted uncertainty spread or uncertainty measures such as quantiles or precentiles (see section **??**).

Because probabilistic forecast solutions have very distinct and different attributes in comparison to deterministic forecast solutions, their testing also needs specific requirements and attention.
The 3 phases (Preparation – During Benchmark/Trial – Post Benchmark/Trial) described in section 3 contain considerations that are equally valid for probabilistic forecast solutions and are recommended to be studies thoroughly before starting a benchmarking or trial process. In the following, specific additional aspects that are recommended to be considered in these three phases will be provided.

4.1 Preparation Phase Challenges for Probabilistic B/T

In the preparation phase of a b/t it is crucial to be aware of that a number of processes that are often applied in a deterministic b/t are not possible or not recommended for probabilistic forecasts.
The most common processes performed by the conductor of a b/t is averaging forecasts and/or aggregating of locations to test, whether, and which combination of forecasts may be better than the best performing forecast.
This is neither a good idea with probabilistic forecasts nor recommended – in some cases it can even lead to wrong results (e.g. aggregating quantiles over locations) – as it undermines to some extent the purpose of probabilistic forecasts. That is, providing a realistic distribution of the uncertainty of the forecast. Also, an average smooth out outliers instead of providing a warning or pinpointing a bad forecast.
A thumb rule may be that, if the task in the b/t is to provide quantiles or percentiles of a specific variable, and it is delivered as multiple or forecast from different vendors, it is **wrong** to:

- aggregate quantiles or percentiles of locations
- average quantiles or percentiles over time

and, although it is theoretically possible (e.g. [13]), it is not recommended to:

- aggregate quantiles or percentiles variables

In contrast to deterministic forecasting, testing probabilistic forecasts does not benefit from aggregation or averaging. In fact, in most cases, it is scientifically not correct to do so. For example, if percentiles are built at multiple locations, aggregating the percentiles and averaging them would lead to a wrong result. In this case, one can only use the raw data of each ensemble member at each location and calculate the percentiles of the aggregated values.
Also, if quantiles have been built from ensemble forecasts from one provider and generated with a statistical approach from another provider, adding quantiles and averaging them would lead to a physically wrong results, because the methodologies computations are fundamentally different (see e.g. sec. 4.3 and 5.2 in Bessa et

al. [1]). Probabilistic forecasts are most useful as a tool to deal with weather and general forecast uncertainties.

To summarise, in our context here, an average or mean of an ensemble forecast, quantile regression forecast or forecasts from other probabilistic methods, even if it represents the underlying uncertainty of the target variable well, would be considered in the same way as a deterministic forecast. And, in that sense, all recommendations from section 3.1, 3.2 and 3.3 would apply. Also, comparisons and averaging with deterministic forecasts may be done and can be expected to score well in general statistical metrics such as MAE, RMSE, BIAS, etc: In other words, such probabilistic forecasts, including minimum, maximum, quantiles or percentiles, can be handled in the same way as deterministic forecasts.

To verify the usefulness, and applicability of a probabilistic forecast to provide specific information about the risk of a certain event to occur or not however, cannot be handled in this way. Here, we need different considerations.

4.2 Evaluation Challenges for probabilistic B/T

Verification scores are useful scientific instruments for the development of forecasts, but often not useful to define the value of a forecast for the end-user. The recommendations we want to make here are therefore following a "forecast value concept, in which ".. forecasts only have value, if a user takes action as a result of a forecast, and that action saves the user money", that was introduced by Mylne [17].

The concept also looks at the importance for the end-user to be able to discriminate uncertainty and spread from forecast scenarios and ensembles, e.g. should I be more or less confident in today's forecast than yesterday's in my decision making.

In this context, it is important to know, how forecasts will be applied in order to find appropriate scoring rules. There is a lot of literature describing statistical metrics and recommendations provided in Part 3 of this recommended practice.

Here, we will provide some recommendations and examples regarding appropriate metrics for typical applications in the context of a trial with one or multiple participants and benchmarking of supplier(s). As described in the key considerations in section 3.1.1 and Part 3 of this recommended practice, it is always a good idea to develop a framework of different metrics and give such different metrics different weights that feed into an overall evaluation score.

The most common and useful scores for probabilistic forecasts used in typical renewable energy applications are:

- **Brier Scores** is the "MAE of probabilistic forecasts" and evaluates categorical forecasts of binary events (see [4, 16, 12])

- **The Continuous Ranked Probability Skill (CRPS) and Energy Score** can be interpreted as the integral of the Brier score over an infinite number of predictand classes of infinitesimal width and with possible threshold values for the parameter under consideration. For a deterministic forecast system, the CRPS reduces to the mean absolute error (see [11, 5, 9, 3])

- **Relative Operating Characteristics (ROC)** measures the skill of a forecast in predicting an event in terms of hit rates and false alarm rates (see [14, 6, 18]).

- **Rank-, Talagrand or PIT[1] histograms** measure the extent to which the spread covers the forecast uncertainty and can reveal BIAS in the probabilistic forecast. A rank histograms does not evaluate resolution – also associated with *sharpness* and *calibration* –, and must be used in conjunction with other forecast tools such as the ROC, Brier scores, or ranked probability scores to generate a more complete picture of the quality of a probabilistic forecast[21, 10, 20].

- **Reliability (Calibration) Diagram** tells how well predicted probabilities of an event correspond to their observed frequencies and provides insight into how well calibrated a probabilistic forecast is and is a complementary metric to the Brier scores (see section 4.3.1 in Part 3 or [4]) and the Relative Operating Characteristics (ROC) curve (see section 4 or [18, 14])

- **Categorial event analysis** in form of e.g. contingency tables or critical success index (CSI) show whether a forecast is targetted towards its purpose, or in other words, what type of errors are being made. With the four categories of hits, misses, false alarm, and correct negatives, it is possible for an end-user to easily associate costs and benefits to a forecast and thereby evaluate its value.

Part 3 of this document series contains a table (Table 5.2) that shows the most common application examples and corresponding recommended evaluation metrics. Details about how to compute or construct the recommended metrics and diagrams can also be found in part 3, section 4.3 of this document series and example code for evaluation, verification and validation can be found in Appendix B of part 3.

[1] probability integral transform

Chapter 5

BEST PRACTICES

Although there are many different ways that a t/b may be conducted, there are some common elements of a successful t/b that provide the t/b operator with the best forecast solution and the participants with useful knowledge of where their forecast ranks among the competition.
The following are some selected best practice recommendations:

1. A clear purpose for the t/b exercise
2. Pre-defined and explicit accuracy metrics and solution selection criteria[1]
3. A clear time line (start/end dates, selection announcement, contract award)
4. Anonymized forecast results. Ask FSP's approval to share results. This helps FSPs find ways to improve their forecast accuracy and see their shortcomings.
5. Question & answer period before benchmark period begins (1-2 weeks)
6. Sufficient time allocated for testing the transfer of data between participant(s) and operator
7. Prompt communication to participants regarding any changes or answers to questions that arise
8. Consistent forecast file format requested of all - example file sent to all
9. Consistent data formats (both observations and forecast files) ideally as close to (if not identical to) what the trial operator needs, once contract is executed.
10. Providing the same historical and project metadata to all participants

[1] See guideline for forecast evaluation and code examples in Part 3 and the reference section 5

11. Allocation of sufficient resources by the t/b conductor to furnish data and perform validation

12. PITFALLS TO AVOID The following list describes a few common mistakes and how to avoid them in the design, setup and execution of a forecast t/b. The consequences of errors and omissions in trials are often underestimated. However, if results are not representative, the efforts that have gone into a t/b can effectively be wasted. Some of these common pitfalls can be expensive to the operator because they result in placing the operator in a position of making a decision without having truly objective and representative information to base it on.

    (a) Poor Communication
    All FSPs should receive the same information. Answers to questions should be shared with all FSPs. Fairness, and perception of fairness, are important when running and evaluating the results of trials.

    (b) Unreliable Validation Results
    Don't compare forecasts from two different power plants or from different time periods. Forecast performance will vary depending on location and specific time periods. Only forecasts for the same period and location/power plant/portfolio should be compared.

    (c) Examples of Bad Design

        i. A trial with 1 month length during a low-wind month
        ii. No on-site observations shared with forecast providers
        iii. Hour-ahead forecasts initiated from once a day data update
        iv. Data only processed in batches or at the end of a real-time trial – this is an invitation for cheating to the FSPs. In most cases, there will be some that use the opportunity to do so

    (d) Examples of Missing or Non-communicated Data

        i. daylight savings time changes are not specified
        ii. data time stamp represents interval beginning or ending not specified
        iii. plant capacity of historical data differs from present capacity
        iv. data about curtailment and maintenance outages not provided

    (e) Possibility of Cheating
    In any type of competition, cheating is a reality. If there are not taken precautions, results may be biased and decisions are taken upon incorrect results. It is recommended that the possibility of cheating is considered with seriousness and avoided, where possible.
    Typical situations, where cheating is being observed are:

(i) Forecast t/b being carried out for a period of time for which FSPs are given data. Recommendation: separate historical data from t/b period.

(ii) if there is one or more incumbent FSP with a longer history of data, this should be taken into consideration in the evaluation, as such an FSP may not be able or willing to modify forecast models for the purpose of being "comparable" in a t/b. Recommendation: see limitations in Table 2 and part 3 of this recommended practice.

(iii) Missing forecasts: FSP leave out "difficult situations" as missing forecasts are often not penalized. However, missing data may bias "average" forecast metrics, potentially resulting in the formulation of incorrect conclusions. Recommendation: remove dates where forecasts are missing for one FSP for all FSPs

(iv) If delivered forecasts from a FSP as part of a live trial are not downloaded, moved or copied in accordance with the operational process being simulated, and certainly before the time period being forecast, FSPs can potentially renew forecasts with high accuracy due to fresher information being available. Recommendation: Such an omission should not be underestimated and care taken for the evaluation.

REFERENCE MATERIAL

Journal and Book Publications

Bessa, R.J.; Möhrlen, C.; Fundel, V.; Siefert, M.; Browell, J.; Haglund El Gaidi, S.; Hodge, B.-M.; Cali, U.; Kariniotakis, G. Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry. Energies, 10, 1402, 2017. Online access: `http://www.mdpi.com/1996-1073/10/9/1402`

Draxl, C., L. K. Berg, L. Bianco, T. A. Bonin, A. Choukulkar, A. Clifton, J. W. Cline, et al. 2019. The Verification and Validation Strategy Within the Second Wind Forecast Improvement Project (WFIP 2). Golden, CO: National Renewable Energy Laboratory. NREL/TP-5000-72553. https://www.nrel.gov/docs/fy20osti/72553.pdf

Lee, Joseph Cheuk Yi and Draxl, Caroline and Berg, Larry K., Evaluating Wind Speed and Power Forecasts for Wind Energy Applications Using an Open-Source and Systematic Validation Framework. Available at SSRN:
https://ssrn.com/abstract=4064500 or http://dx.doi.org/10.2139/ssrn.4064500
J. Kehler and D. McCrank, Integration of wind power into Albertas electric system and market operation, Proc. of IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, pp. 1-6. doi: 10.1109/PES.2008.4596824, 2008.

E. Lannoye, A. Tuohy, J. Sharp, V. Von Schamm, W. Callender, L.Aguirre, Solar Power Forecasting Trials and Trial Design: Experience from Texas, Proc. of 5th International Workshop on the Integration of Solar Power into Power Systems,, Brussels, Belgium, ISBN: 978-3-9816549-2-9, 2016.

E. Lannoye, A Tuohy, J Sharp, and W Hobbs, Anonymous Solar Forecasting Trial Outcomes, Lessons learned and trial recommendations, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017.

C. Möhrlen, C. Collier , J. Zack , J. Lerner , Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017. Online access: http://download.weprog.com/WIW2017-292_moehrlen_et-al_v1.pdf.

Conference Papers

Corinna Möhrlen, Recommended Practices for the Implementation of Wind Power Forecasting Solutions Part 1: Forecast Solution Selection Process, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018. Online Access: http://www.ieawindpowerforecasting.dk/publications

Corinna Möhrlen, John Zack, Jeff Lerner, Aidan Tuohy, Jethro Browell, Jakob W. Messner, Craig Collier, Gregor Giebel
Part 2&3: DESIGNING AND EXECUTING FORECASTING BENCHMARKS AND TRIALS AND EVALUATION OF FORECAST SOLUTIONS, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018
Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen, R. Bessa, Understanding Uncertainty: the difficult move from a deterministic to a probabilistic world, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018
Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen, R. Bessa, G. Giebel, J. Jørgensen,G. Giebel, Uncertainty Forecasting Practices for the Next Generation Power System, Proc. 16th Int. Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Berlin (DE), 26-29 June 2017.
Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen (WEPROG, Denmark), C. Collier (DNV GL, USA), J. Zack (AWS Truepower, USA), J. Lerner (Vaisala, USA) Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. 16th Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission

Networks for Offshore Wind Power Plant , Berlin (DE), 26-29 June 2017.
Online Access: http://www.ieawindpowerforecasting.dk/publications

Presentations

C. Collier (2017), Why Do Forecast Trials Often Fail to Answer the Questions for which End-Users Need Answers: A Forecaster's Point of View UVIG Forecasting Workshop, Atlanta (US), 21-22 June 2017. Online access: http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_ccollier.ashx?la=da

T. Maupin (2017), Wind and Solar Forecasting Trials: Do's and Don'ts, Part 1 Best practices. UVIG 2017 Forecasting Workshop, Atlanta (US), 21-22 June 2017. Online access: http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_ccollier.ashx?la=da

C. Möhrlen, C. Collier , J. Zack , J. Lerner , Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017. Online access: http://download.weprog.com/WIW17-292_MOEHRLEN-ET-AL_PRESENTATION_20171028.pdf

J. W. Zack (2017), Wind and solar forecasting trials experience: do's and don'ts, Part 2 UVIG 2017 Forecasting Workshop, Atlanta (US), 21-22 June 2017
Online access: http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_jzack.ashx?la=da

Example Validation and Verification Code Projects

WEvalidate: Python-based code base as a platform to consistently evaluate wind-power forecasts. The tool WE-Validate is meant to gear towards forecast validation using observations and simulations for wind energy ("WE") applications. This infrastructure code enables comparison of time series from arbitrary data sources using user-defined metrics. The tool is designed to be simple, readily usable, open source, publicly available, modularized, and extensible by users. We have detailed instructions for users on its GitHub page, Accessible Online: https://github.com/joejoeyjoseph/WE-Validate

WE-verify-prob: R-based example code base to verify probabilistic wind energy forecasts. The tool WE-verify-prob is a project initiated within the IEA Wind Task 36 and 51 in order to provide example code to the IEA Wind Recommended Practice for the Implemention of Renewable Energy Forecast Solutions Part 2 *Designing and Executing Forecasting Benchmarks and Trials* and 3 *Forecast Solution Evaluation*. Accessible for download at the IEA Wind Task 36 webpage `https://iea-wind.org/task-36/task-36-publications/recommended-practice/`

Code Examples from related projects with relevance to recommendations

The following selection of VV code examples links from related projects with relevance to the recommendations made in this document.

Weather Forecast Verification Utilities: The R-Package 'verification' [8] has been developed for "verifying discrete, continuous and probabilistic forecasts, and forecasts expressed as parametric distributions" by Eric Gilleland from NCAR Research Applications Laboratory[8]. The Package contains all relevant metrics described in chapter **??** and chapter **??**, specifically those described in section **??** and those described by [7]. It can be accessed via CRAN `https://cloud.r-project.org/web/packages/verification/index.html`.

Forecast Verification Routines for Ensemble Forecasts of Weather and Climate: The R-package 'SpecsVerification' [19] is a collection of forecast verification routines developed for the SPECS FP7 project. The emphasis is on comparative verification of ensemble forecasts of weather and climate. The package contains most of the metrics described in chapter **??** and chapter **??**, specifically those described in section **??** and those described by [7]. It can be accessed via CRAN `https://CRAN.R-project.org/package=SpecsVerification`.

Ensemble Forecast Verification for Large Data Sets: The R-package 'easyVerification' [2] is a set of tools to simplify application of forecast verification metrics for (comparative) verification of ensemble forecasts to large data sets. The forecast metrics are partially imported from the 'SpecsVerification' R-package, with additional forecast metrics provided within this package. New user-defined forecast scores can be implemented using the example scores provided and applied using the functionality of this package. The package contains all of the metrics described in chapter **??** and chapter **??**, specifically those described in section **??** and those described by [7]. It can be accessed via CRAN `https://CRAN.R-project.org/`

`package=easyVerification`.

Ensemble Postprocessing with R: The R-package "ensemblepp" [15] provides postporcessing and verification Data Sets and code examples for the chapter "Ensemble Postprocessing with R" of the book "Statistical Postprocessing of Ensemble Forecasts" by Stephane Vannitsem, Daniel S. Wilks, and Jakob W. Messner (2018), Elsevier, 362pp. These data sets contain temperature and precipitation ensemble weather forecasts and corresponding observations at Innsbruck/Austria. Additionally, a demo with the full code of the book chapter is provided. Evaluation code is provided as scatter plots, rank histogram, spread skill relationship and histograms. Available Version: 1.0-0 online: `https://cran.r-project.org/web/packages/ensemblepp`.

GLOSSARY AND ABBREVIATIONS

Ensemble Forecasting:
Ensemble forecasts are sets of different forecast scenarios, which provide an objective way of evaluating the range of possibilities and probabilities in a (weather or weather related) forecast.

Probabilistic Forecast:
General description of defining the uncertainty of a forecast with objective methods. These can be ensemble forecasts, probability of exceedance forecasts, or other forms of measures of uncertainty derived by statistical models.

Quantile:
A quantile is the value below which the observations/forecasts fall with a certain probability when divided into equal-sized, adjacent, subgroups.

Quartile:
quantiles that divide the distribution into four equal parts.

Percentile:
Percentiles are quantiles where this probability is given as a percentage (0-100) rather than a number between 0 and 1.

Decile:
Quantiles that divide a distribution into 10 equal parts.

Median:
the 2nd quantile, 50th percentile or 5th decile, i.e. the value, where the distribution has equally many values above and below that value.

Abbreviations

The following abbreviations are used in this document:

FSP	Forecast service provider
NWP	Numerical Weather Prediction
EPS	Ensemble Prediction System
NDA	Non-disclosure Agreement
RFI	Request for Information
RFP	Request for Proposals
TSO	Transmission system operators
ISO	Independent system operator

Bibliography

[1] Botterud A Wang J Bessa RJ Miranda V. "'Good' or 'bad' wind power forecasts: a relative concept". In: *Wind Energy* 14.5 (2010), 625˘636. DOI: 10.1002/we.444.

[2] Jonas Bhend et al. *easyVerification: Ensemble Forecast Verification for Large Data Sets*. Version 0.4.4. R (3.0), SpecsVerification (0.5), stats, utils. 2020. URL: https://CRAN.R-project.org/package=easyVerification.

[3] Mathias Blicher Bjerregård, Jan Kloppenborg Møller, and Henrik Madsen. "An introduction to multivariate probabilistic forecast evaluation". In: *Energy and AI* 4 (2021), p. 100058. ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2021.100058. URL: https://www.sciencedirect.com/science/article/pii/S2666546821000124.

[4] GLENN W. BRIER. "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY". In: *Monthly Weather Review* 78.1 (1950), pp. 1 –3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[5] G. Candille and O. Talagrand. "Evaluation of probabilistic prediction systems for a scalar variable". In: *Quarterly Journal of the Royal Meteorological Society* 131.609 (2005), pp. 2131–2150. DOI: https://doi.org/10.1256/qj.04.71. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.71. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.04.71.

[6] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[7] WWRP/WGNE Joint Working Group on Forecast Verification Research. In: Berlin, Germany. URL: http://www.cawcr.gov.au/projects/verification/.

[8] Eric Gilleland. *verification: Weather Forecast Verification Utilities*. R package version 2.10), methods, fields, boot, CircStats, MASS, dtw. 2015. URL: https://CRAN.R-project.org/package=verification.

[9] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (2007), pp. 243–268. DOI: https://doi.org/10.1111/j.1467-9868.2007.00587.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00587.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x.

[10] Thomas M. Hamill. "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". In: *Monthly Weather Review* 129.3 (2001), pp. 550 –560. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

[11] Hans Hersbach. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems". In: *Weather and Forecasting* 15.5 (2000), pp. 559 –570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

[12] Ian T. Jolliffe and David B. Stephenson. "Proper Scores for Probability Forecasts Can Never Be Equitable". In: *Monthly Weather Review* 136.4 (2008), pp. 1505 –1510. DOI: 10.1175/2007MWR2194.1. URL: https://journals.ametsoc.org/view/journals/mwre/136/4/2007mwr2194.1.xml.

[13] Kenneth Jr, Yael Grushka-Cockayne, and Robert Winkler. "Is It Better to Average Probabilities or Quantiles?" In: *Management Science* 59 (Mar. 2012). DOI: 10.2139/ssrn.2066806.

[14] I. Mason. "A model for assessment of weather forecasts". In: *Australian Meteorological Magazin* 30 (1982), pp. 291–303.

[15] Jakob W. Messner. *ensemblepp: Ensemble Postprocessing Data Sets*. Version 1.0-0.

[16] Allan H. Murphy. "A New Vector Partition of the Probability Score". In: *Journal of Applied Meteorology and Climatology* 12.4 (1973), pp. 595 –600. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

[17] Kenneth R. Mylne. "Decision-making from probability forecasts based on forecast value". In: *Meteorological Applications* 9.3 (2002), pp. 307–315. DOI: https://doi.org/10.1017/S1350482702003043. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1017/S1350482702003043. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482702003043.

[18] *Receiver operating characteristic*. https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed: 2021-10-11.

[19] Stefan Siegert et al. *SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate*. Version 0.5-3. License GPL-2 | GPL-3 [expanded from: GPL (2)]. 2020. URL: https://CRAN.R-project.org/package=verification.

[20] Leonard A. Smith and James A. Hansen. "Extending the Limits of Ensemble Forecast Verification with the Minimum Spanning Tree". In: *Monthly Weather Review* 132.6 (2004), pp. 1522 –1528. DOI: 10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/132/6/1520-0493_2004_132_1522_etloef_2.0.co_2.xml.

[21] O. Talagrand, R. Vautard, and B. Strauss. "Evaluation of probabilistic prediction systems". In: Shinfield Park, Reading, UK, 1999.

Bibliography

[1] Botterud A Wang J Bessa RJ Miranda V. “’Good’ or ’bad’ wind power forecasts: a relative concept”. In: *Wind Energy* 14.5 (2010), 625˘636. DOI: 10.1002/we.444.

[2] Jonas Bhend et al. *easyVerification: Ensemble Forecast Verification for Large Data Sets*. Version 0.4.4. R (3.0), SpecsVerification (0.5), stats, utils. 2020. URL: https://CRAN.R-project.org/package=easyVerification.

[3] Mathias Blicher Bjerregård, Jan Kloppenborg Møller, and Henrik Madsen. “An introduction to multivariate probabilistic forecast evaluation”. In: *Energy and AI* 4 (2021), p. 100058. ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2021.100058. URL: https://www.sciencedirect.com/science/article/pii/S2666546821000124.

[4] GLENN W. BRIER. “VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY”. In: *Monthly Weather Review* 78.1 (1950), pp. 1 –3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[5] G. Candille and O. Talagrand. “Evaluation of probabilistic prediction systems for a scalar variable”. In: *Quarterly Journal of the Royal Meteorological Society* 131.609 (2005), pp. 2131–2150. DOI: https://doi.org/10.1256/qj.04.71. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.71. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.04.71.

[6] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[7] WWRP/WGNE Joint Working Group on Forecast Verification Research. In: Berlin, Germany. URL: http://www.cawcr.gov.au/projects/verification/.

[8] Eric Gilleland. *verification: Weather Forecast Verification Utilities*. R package version 2.10), methods, fields, boot, CircStats, MASS, dtw. 2015. URL: https://CRAN.R-project.org/package=verification.

[9] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (2007), pp. 243–268. DOI: https://doi.org/10.1111/j.1467-9868.2007.00587.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00587.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x.

[10] Thomas M. Hamill. "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". In: *Monthly Weather Review* 129.3 (2001), pp. 550 –560. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

[11] Hans Hersbach. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems". In: *Weather and Forecasting* 15.5 (2000), pp. 559 –570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

[12] Ian T. Jolliffe and David B. Stephenson. "Proper Scores for Probability Forecasts Can Never Be Equitable". In: *Monthly Weather Review* 136.4 (2008), pp. 1505 –1510. DOI: 10.1175/2007MWR2194.1. URL: https://journals.ametsoc.org/view/journals/mwre/136/4/2007mwr2194.1.xml.

[13] Kenneth Jr, Yael Grushka-Cockayne, and Robert Winkler. "Is It Better to Average Probabilities or Quantiles?" In: *Management Science* 59 (Mar. 2012). DOI: 10.2139/ssrn.2066806.

[14] I. Mason. "A model for assessment of weather forecasts". In: *Australian Meteorological Magazin* 30 (1982), pp. 291–303.

[15] Jakob W. Messner. *ensemblepp: Ensemble Postprocessing Data Sets*. Version 1.0-0.

[16] Allan H. Murphy. "A New Vector Partition of the Probability Score". In: *Journal of Applied Meteorology and Climatology* 12.4 (1973), pp. 595 –600. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

[17] Kenneth R. Mylne. “Decision-making from probability forecasts based on forecast value”. In: *Meteorological Applications* 9.3 (2002), pp. 307–315. DOI: https://doi.org/10.1017/S1350482702003043. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1017/S1350482702003043. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482702003043.

[18] *Receiver operating characteristic*. https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed: 2021-10-11.

[19] Stefan Siegert et al. *SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate*. Version 0.5-3. License GPL-2 | GPL-3 [expanded from: GPL (2)]. 2020. URL: https://CRAN.R-project.org/package=verification.

[20] Leonard A. Smith and James A. Hansen. “Extending the Limits of Ensemble Forecast Verification with the Minimum Spanning Tree”. In: *Monthly Weather Review* 132.6 (2004), pp. 1522 –1528. DOI: 10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/132/6/1520-0493_2004_132_1522_etloef_2.0.co_2.xml.

[21] O. Talagrand, R. Vautard, and B. Strauss. “Evaluation of probabilistic prediction systems”. In: Shinfield Park, Reading, UK, 1999.

Appendix A

Metadata Checklist

The following checklist (Table A.1), when filled out, will greatly aid FSPs in configuring forecasts efficiently. Many of the essential questions relevant to benchmark and trial forecast model configuration are provided here.

Note that the following table is an example and may not contain all necessary information required for the FSP to setup a solution for your purpose. The table is meant to serve as a guideline and can be copied, but should be carefully adopted to the specific exercises before sending out to FSP with questions filled in. If this is done with care, it will expedite forecast configuration and save back and forth communication time.

Table A.1: Wind Power Forecast Trial Checklist

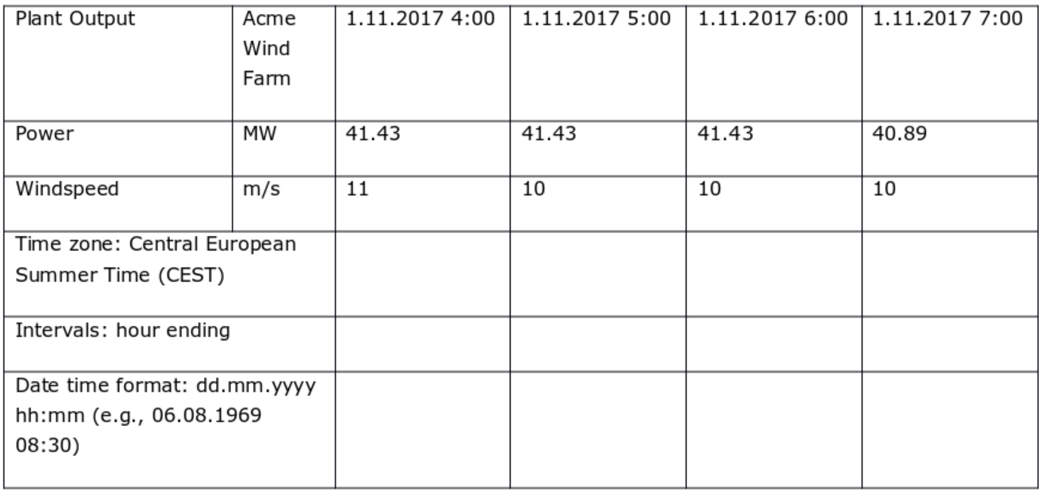
Metadata	**Input**
Name of site(s) as it should appear in datafile	
Name of site(s) as it should appear in datafile	
Latitude and longitude coordinates of sites	
Nameplate capacity of each site	
Will a graphical web tool be needed?	
Turbine make/model/rating	
Number of turbines	
Hub height of turbines	
Please attach suitable plant power curve	

Metadata	**Input**
Forecast output information	
Forecast output time intervals (e.g., 15-min, 1-hourly)	
Length of forecast required	
Timezone of forecast datafile	
Will local daylight savings time be needed?	
Forecast update frequency (e.g., once a day, every hour)	
Value of Forecast	
Which variables will be forecasted and validated?	
Which forecast horizons are being validated?	
Which metrics are being used to gage forecast performance?	
List criteria for determining winning forecast provider	
Will results be shared as a report? Will results be anonymized?	
On what frequency will results be shared with forecast provider?	
Historical Data Checklist	
Is the data in UTC or local time?	
Is the data interval beginning or ending or instantaneous?	
What are the units of the data?	
If met tower histories being provided, indicate height of measurements.	
Realtime Data Checklist (if applicable)	
Is the data in UTC or local time?	
Is the data interval beginning or ending or instantaneous?	
What are the units of the data?	
Email and Telephone number of technical point of contact (POC)	
Email and Telephone of datafeed POC	
Name and email of users that need website access	
Person name and email that filled out this checklist	

Appendix B

Sample forecast file structures

Back and forth communication can sometimes delay the start of a trial or benchmark. One of these delays is getting the forecast file output format just right for the beginning of the trial.
Standardisation of the format will make the trial operators life much easier when time comes to validating forecasts. A best practice here is for the trial operator to use a format that is already in use or a format that has already proven to work in operations.



Plant Output	Acme Wind Farm	1.11.2017 4:00	1.11.2017 5:00	1.11.2017 6:00	1.11.2017 7:00
Power	MW	41.43	41.43	41.43	40.89
Windspeed	m/s	11	10	10	10
Time zone: Central European Summer Time (CEST)					
Intervals: hour ending					
Date time format: dd.mm.yyyy hh:mm (e.g., 06.08.1969 08:30)					

Figure B.1: Example forecast file with the first few fields.

B.0.1 XSD template example for forecasts and SCADA

The following are typical XSDs for forecasts and SCADA data in a b/t, usable also with WebServices

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="WindForecast">
    <xs:complexType>
     <xs:attribute name="VendorCode" type="xs:string" use="required"/>
     <xs:attribute name="ImportTime" type="xs:dateTime" use="required"/>
     <xs:sequence>
       <xs:element name="CUSTOMER">
         <xs:complexType>
          <xs:attribute name="name" type="xs:string" use="required"/>
           <xs:sequence>
            <xs:element name="Forecast">
              <xs:complexType>
               <xs:attribute name="MWaggregated" type="xs:double" use="required"/>
               <xs:attribute name="time" type="xs:dateTime" use="required" />
               <xs:sequence>
                 <xs:element name="Probability">
                   <xs:complexType>
                     <xs:attribute name="P95" type="xs:double" use="required"/>
                     <xs:attribute name="P50" type="xs:double" use="required"/>
                     <xs:attribute name="P05" type="xs:double" use="required"/>
                     <xs:attribute name="max" type="xs:double" use="required"/>
                     <xs:attribute name="min" type="xs:double" use="required"/>
                    </xs:complexType>
                  </xs:element>
                 <xs:element name="WindFarms">
                 <xs:complexType>
               <xs:sequence>
                 <xs:element name="WindPark1">
                  <xs:complexType>
                      <xs:attribute name="id" type="xs:string" use="required"/>
                       <xs:attribute name="mw" type="xs:double" use="required"/>
                   </xs:complexType>
                   </xs:element>
               </xs:sequence>
                   </xs:complexType>
                   </xs:element>
               </xs:sequence>
                </xs:complexType>
              </xs:element>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

B.0.2 XSD SCADA template for exchange of real-time measurements

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="WindSCADA">
  <xs:complexType>
    <xs:sequence>
      <xs:element maxOccurs="unbounded" name="WindPark">
        <xs:complexType>
          <xs:attribute name="ID" type="xs:string" use="required"/>
          <xs:attribute name="Time" type="xs:dateTime" use="required"/>
          <xs:attribute name="Mw" type="xs:decimal" use="required"/>
          <xs:attribute name="Availabilty" type="xs:decimal" use=" optional"/>
          <xs:attribute name="CurrentActivePower" type="xs:decimal" use=" optional"/>
          <xs:attribute name="Curtailment" type="xs:string" use="optional"/>
          <xs:attribute name="WindSpeed" type="xs:decimal" use="optional"/>
          <xs:attribute name="WindDirection" type="xs:decimal" use="optional"/>
          <xs:attribute name="AirTemperature" type="xs:decimal" use="optional"/>
          <xs:attribute name="AirPressure" type="xs:decimal" use="optional"/>
        <xs:attribute name="Outage" type="xs:decimal" use="optional"/>
      </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
</xs:schema>
```