

|  |  |  |
| --- | --- | --- |
| PNNL-XXXXX |  | |
|  | Evaluating wind speed and power forecasts for wind energy applications using an open-source and systematic validation framework  January 2022  Joseph C Y Lee1  Caroline Draxl2  Larry K Berg1  William J Shaw1  1 Pacific Northwest National Laboratory  2 National Renewable Energy Laboratory | |
|  |  |  |
| Prepared for the U.S. Department of Energy  under Contract DE-AC05-76RL01830 |

**DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY

*operated by*

BATTELLE

*for the*

UNITED STATES DEPARTMENT OF ENERGY

*under Contract DE-AC05-76RL01830*

**Printed in the United States of America**

**Available to DOE and DOE contractors from the**

**Office of Scientific and Technical Information,**

**P.O. Box 62, Oak Ridge, TN 37831-0062;**

**ph: (865) 576-8401**

**fax: (865) 576-5728**

**email:** [**reports@adonis.osti.gov**](mailto:reports@adonis.osti.gov)

**Available to the public from the National Technical Information Service**

**5301 Shawnee Rd., Alexandria, VA 22312**

**ph: (800) 553-NTIS (6847)**

**email: orders@ntis.gov <**[**https://www.ntis.gov/about**](https://www.ntis.gov/about)**>**

**Online ordering:** [**http://www.ntis.gov**](http://www.ntis.gov)

Evaluating wind speed and power forecasts for wind energy applications using an open-source and systematic validation framework

January 2022

Joseph C Y Lee1

Larry K Berg1

Caroline Draxl2

William J Shaw1

1 Pacific Northwest National Laboratory

2 National Renewable Energy Laboratory

Prepared for  
the U.S. Department of Energy  
under Contract DE‑AC05‑76RL01830

Pacific Northwest National Laboratory

Richland, Washington 99354

Abstract

Building on the verification and validation work developed under the Second Wind Forecast Improvement Project, this work exhibits the value of a consistent procedure to evaluate wind power forecasts. We established an open-source Python code base tailored for wind speed and wind power forecast validation, WE-Validate. The code base can evaluate model forecasts with observations in a coherent manner. To demonstrate the systematic validation framework of WE-Validate, we designed and hosted a forecast evaluation benchmark exercise. We invited forecast providers in industry and academia to participate and submit forecasts for two case studies. We then evaluated the submissions with WE-Validate. Our findings suggest that ensemble means have reasonable skills in time-series forecasting, whereas they are often inferior to single ensemble members in wind ramp forecasting. Adopting a voting scheme in ramp forecasting that allows ensemble members to detect ramps independently leads to satisfactory skill scores. Throughout this document, we also emphasize the importance of using statistically robust and resistant metrics as well as equitable skill scores in forecast evaluation.

Acknowledgments

We thank Markus Kreklau and his colleagues at the Bundesamt für Seeschifffahrt und Hydrographie for providing the meteorological tower observations for the FINO2 case study. We thank Wagner Uwe and his colleagues at Energie Baden-Württemberg AG for providing the Baltic-2 wind farm measurements. We also thank Vaisala for providing the data of the Vaisala Triton Sodar Wind Profiler for the Second Wind Forecast Improvement Project case study, and the data are publicly available at the Atmosphere to Electrons Data Archive and Portal (<https://a2e.energy.gov/data>).

We are grateful to our collaborators who submitted data for the benchmark exercise. We thank the members of the International Energy Agency Wind Task 36 for their valuable input during the conception phase of the benchmark exercise and WE-Validate. Special thanks go to Helmut Frank and his colleagues at Deutscher Wetterdienst; John Zack and his colleagues at MESO, Inc.; Jana Fishereit, Gregor Giebel, Marc Imberger, Xiaoli Guo Larsén, and Andrea Hahmann at the Department of Wind Energy at the Technical University of Denmark; and Corinna Möhrlen and her colleagues at WEPROG.

This work was authored by Pacific Northwest National Laboratory, operated for the U.S. Department of Energy by Battelle under contract DE-AC05-76RL01830.

This work was authored [in part] by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Acronyms and Abbreviations

|  |  |
| --- | --- |
| API | application programming interface |
| CSI | critical success index |
| DOE | U.S. Department of Energy |
| FINO2 | Forschungsplattformen in Nord- und Ostsee Nr. 2 |
| FN | false negative |
| FP | false positive |
| IEA | International Energy Agency |
| POD | probability of detection |
| PSS | peirce skill score |
| RMSE | root-mean-square error |
| SEDS | symmetric extreme dependency score |
| SR | success ratio |
| TCP | Technical Collaboration Programme |
| TN | true negative |
| TP | true positive |
| V&V | verification and validation |
| WFIP2 | The Second Wind Forecast Improvement Project |

**Contents**

[Abstract ii](#_Toc94021863)

[Acknowledgments iii](#_Toc94021864)

[Acronyms and Abbreviations iv](#_Toc94021865)

[Contents v](#_Toc94021866)

[1.0 Motivation 1](#_Toc94021867)

[2.0 WE-Validate 2](#_Toc94021868)

[3.0 Metrics evaluation 4](#_Toc94021869)

[3.1 Single-value metrics for time series forecasts 4](#_Toc94021870)

[3.2 Metrics for ramp forecasts 5](#_Toc94021871)

[4.0 Benchmark exercise 8](#_Toc94021872)

[5.0 Analysis of benchmark submissions 12](#_Toc94021873)

[5.1 Time series forecasts 12](#_Toc94021874)

[5.2 Ramp forecasts 18](#_Toc94021875)

[6.0 Discussions 29](#_Toc94021876)

[7.0 Conclusions 31](#_Toc94021877)

[8.0 References 32](#_Toc94021878)

Figures

[Figure 1. The geographical location of the WFIP2 sodar in the Pacific Northwest region of the United States. 10](#_Toc94021879)

[Figure 2. The geographical locations of the FINO2 tower and the Baltic-2 wind farm in the North Sea of Europe. Because we could not share the specific turbine locations with the participants, we asked them to submit spatially averaged Baltic-2 forecasts in the area bounded by the black lines. 10](#_Toc94021880)

[Figure 3. (top) Wind speed time series of the WFIP2 case at 80 m above ground level, where the red line illustrates cup anemometer measurements, the magenta line denotes the p1 forecast, the purple, light blue, and cyan lines, respectively, indicate ensemble mean forecasts from p2, p4, and p5, and the black line is the ensemble average of the four participants. (middle) Similar to (top), but for wind speed of the FINO2 case at 82 m above sea level with submissions from five participants. (bottom) Similar to (middle), but for wind-farm power data at Baltic-2 at 78.25 m above sea level. Across the panels, the shading around the ensemble means of p4 and p5 represents the standard deviation of the ensemble members around the mean. The ensemble means of p2 and p3 are plotted as single lines and no shading is incorporated because the differences between the two ensemble members are mostly trivial. 13](#_Toc94021881)

[Figure 4. Swarm plots of root-mean-square error (left) and median absolute error (right) of the p4 ensemble (blue), the p5 ensemble (cyan), and the 5-organization ensemble (gray) on the forecasts of WFIP2 wind speeds at 80 m above ground level (top), FINO2 wind speeds at 82 m above sea level (middle), and Baltic-2 hub height power (bottom). In each data column, each dot represents an ensemble member, and the cross indicates the respective ensemble average. 14](#_Toc94021882)

[Figure 5. The rank of the absolute error of hourly power forecast among six time series: p1, the four intra-organization ensemble means, and the 5-organization ensemble mean (top left as “avg”) at the Baltic-2 wind farm. A member is ranked 1 for an hour when its power forecast error is the lowest among all members, including the ensemble mean, at that hour. Each pie chart illustrates the portion of the ranks over the 7-day period that each member holds. 15](#_Toc94021883)

[Figure 6. Histograms of absolute errors of hourly power forecast over the 7-day period at the Baltic-2 wind farm. The 5-organization ensemble mean is marked as “avg.” 16](#_Toc94021884)

[Figure 7. Power spectra of the hourly power forecast error of the Baltic-2 case over the 7-day period for the p4 ensemble (top), the p5 ensemble (bottom), and the 5-organization ensemble (bottom). In each spectrum, the black line indicates the ensemble mean and the other colored lines represent the ensemble members. 17](#_Toc94021885)

[Figure 8. Illustration of the 2x2 contingency table on the power ramp forecast at the Baltic-2 wind farm over the 7-day period. Each row represents a ramp event definition, from changing over |50 MW| within 2 hours to changing over |50 MW| within 8 hours, and the ramp definitions include both up ramps and down ramps. Each column is an ensemble mean, except for p1, which submitted single-member forecasts. The four triangles in each square characterize the 2x2 contingency table, as percentages of instances among the four parameters. The percentages are annotated in the triangles. The sum of the four triangles in each square is 100%. 19](#_Toc94021886)

[Figure 9. Illustration of determining wind power ramps with voting scheme using a ramp definition of |50 MW| within 6 hours for the Baltic-2 case. In this example, two out of five voting members indicate ramps between 6 UTC and 12 UTC on 5 October, so the 40% voting scheme labels ramp in this period (gray). Similarly, five out of five voting members indicate ramps between 17 UTC and 23 UTC on 5 October, and three out of five voting members indicate ramps between 4 UTC and 10 UTC on 6 October, so the respective 100% (yellow) and 60% (green) voting schemes label ramps for the two periods. 20](#_Toc94021887)

[Figure 10. The counts of true positive wind power ramp forecasts over the 7-day period at the Baltic-2 wind farm. Each row represents a ramp event definition, and all ramp definitions include both up ramps and down ramps. From left to right, the columns illustrate different voting schemes (XX% vote), the 5-organization ensemble mean (all 5p mean), and the forecasts from each organization (pX or pX mean). 21](#_Toc94021888)

[Figure 11. Similar to Figure 10, but for false positive forecasts. 22](#_Toc94021889)

[Figure 12. Similar to Figure 10, but for false negative forecasts. 23](#_Toc94021890)

[Figure 13. Similar to Figure 10, but for true negative forecasts. 24](#_Toc94021891)

[Figure 14. Similar to Figure 10, but for the Peirce skill score using the 2x2 contingency table projected in Figures 10, 11, 12, and 13. 25](#_Toc94021892)

[Figure 15. Similar to Figure 10, but for the symmetric extreme dependency score using the 2x2 contingency table projected in Figures 10, 11, 12, and 13. A blank cell means that the SEDS is undefined, where the forecast does not yield any TP forecasts under a given ramp definition. 26](#_Toc94021893)

[Figure 16. Performance diagrams of the Baltic-2 power-ramp forecasts for the 7-day period under the ramp definition of changing over |50 MW| within 6 hours. The x-axis is success ratio, the y-axis is probability of detection, the dashed diagonal lines extending from the origin represent bias values, and the blue contours indicate a range of critical success index values. A perfect forecast would land on the top-right corner of the diagram. The top diagram depicts the ensemble means of each participant as purple dots, the results of voting schemes as green triangles, and the 5-organization ensemble mean as a black cross. The bottom diagram illustrates the p5 ensemble members as gray dots, the results of voting schemes among the p5 ensemble members as triangles, and the p5 ensemble mean as a black cross. 28](#_Toc94021894)

Tables

[Table 1. Summary of the two cases 8](#_Toc94021895)

[Table 2. Summary of collected forecast submissions 11](#_Toc94021896)

[Table 3. Observed ramp counts with different ramp definitions at the Baltic-2 wind farm during the 7-day period 18](#_Toc94021897)

# Motivation

Selecting accurate renewable energy forecasts that suit one’s needs requires careful assessment. For instance, variations emerge among forecast providers, from numerical modeling practices to communicating forecast uncertainties (Bessa et al. 2017). Varying methodologies chosen by different organizations also create uncertainties in predicting wind energy production (Lee and Fields 2021; Craig et al. 2018). Creating a benchmark to evaluate forecast performance is also costly for forecast users and providers (Möhrlen and Zack 2019). Therefore, to minimize the misalignment of expectations and requirements on wind energy forecasts among stakeholders, a comprehensive and objective process of selecting forecast providers has been proposed (Möhrlen and Zack 2019; Möhrlen, Zack, and Lerner 2019).

Adhering to the recommended practices, in this work we illustrate a systematic approach to evaluate forecast model performance with observations. We developed an open-source code base for wind power forecast validation, WE-Validate, that solidifies the rigorous forecast evaluation framework. This framework allows for transparency in model evaluation, provides clear guidance in operational settings, and enables new research endeavors.

We hosted a forecast benchmark exercise that involved industry and academia collaboration. With WE-Validate, we established a systematic forecast validation framework with the ability to coherently evaluate multiple forecasts within and across various organizations, with an emphasis on evaluating forecasts of wind ramps. Users of WE-Validate can provide forecast and observation time series and evaluate model performance in a manner consistent with others who use the tool. In this work, we exhibit the results of the benchmark exercise using WE-Validate, discuss the characteristics of different evaluation metrics, and reveal the strengths and weaknesses of ensemble mean forecasts.

This study was funded and carried out as an extension of the model verification and validation (V&V) effort (Draxl et al. 2019) of the Second Wind Forecast Improvement Project (WFIP2) (Wilczak et al. 2019; Shaw et al. 2019; Olson et al. 2019) through the U.S. Department of Energy. In addition to being an extension of the WFIP2 work, this study also represents a contribution to Phase II of Task 36: Wind Power Forecasting of the International Energy Agency’s (IEA’s) Wind Technical Collaboration Programme. The goal of Task 36 is to improve the value of wind energy forecasts to the wind industry (Giebel et al. 2018). This work falls under the umbrella of Work Package 1 within IEA Wind Task 36, which focuses on forecast model improvement. By being part of IEA Wind Task 36, this study benefitted from the invaluable input from task members, and the application of WE-Validate can be directly linked to real-world applications. In particular, the Recommended Practices for Selecting Renewable Power Forecasting Solutions (Möhrlen and Zack 2019; Möhrlen, Zack, and Lerner 2019) were consulted during the design and conception phase of the benchmark exercise and code development.

# WE-Validate

We developed a Python-based code base as a platform to consistently evaluate wind power forecasts. We call our tool WE-Validate to gear toward forecast validation using observations and simulations for wind energy (WE) applications. This infrastructure code enables the comparison of time series from arbitrary data sources using user-defined metrics. This 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, [github.com/joejoeyjoseph/WE-Validate](https://github.com/joejoeyjoseph/WE-Validate). The tool is currently tailored for wind power forecast evaluation, and it can be extended to solar forecasting and other applications as well.

The tool was built on the data structure of pandas (Reback et al. 2020), which is a widely used Python package. The tool has built-in data quality control capabilities, such as checking, flagging, and removing missing or duplicated data; aligning multiple time series to user-defined start and end times of the evaluation period; resampling higher-frequency data to match another data set of coarser resolution.

The code can handle data inputs at various height levels and data frequencies. After the initial data quality control steps, at each user-defined height, the code compares the observed time series to the modeled time series and computes the evaluation metrics (Section 3.1). If multiple forecasts are specified (e.g., ensemble forecasts), the code would screen all the individual forecasts and compare each of them with the observations. Instead of using observations, WE-Validate users can also employ time series from a reference simulation and compare it against other simulated time series to examine model differences and improvements.

For visualization, the code generates a time series line plot, a histogram, and a scatterplot between the forecast and observed values at each specified height. When ramp evaluation is turned on, for each ramp definition, the code computes the ramp evaluation metrics (Section 3.2) and generate a time series line plot overlayed with a 2x2 contingency table. When the variable of interest is wind speed and a height level matches the specified hub height, the code derives power using a power curve, which can be selected by the user.

Users edit a configuration file, which is in yaml format, and execute the tool in Python. The configuration file specifies the details of the forecast evaluation, such as the period of evaluation, the heights of interest, the variables to evaluate, the forecast and observed data frequency, the method of time step alignment, and the ramp definitions. An existing code example listed on GitHub uses a Jupyter Notebook, but executing the code does not require using one.

In a configuration file, when the user specifies the same data processing settings to different forecast data sets, they can consistently compare the resultant evaluation metrics from across the data sets. For instance, if a wind farm operator receives a forecast at 10-minute resolution and another forecast at 30-minute resolution, the analyst can execute WE-Validate at a 30-minute resolution for both forecasts and compare their performance in a compatible manner.

We note that other validation tools also exist or are under development. The WindSider ([windsider](http://www.windsider.io).io) is tailored for the wind resource assessment process and uses the data structure of xarray, another Python package. As of this writing, the WindSider is under development, led by experts from 3E and the Technical University of Denmark. The Ramp Tool and Metric (Bianco et al. 2016), created by researchers at the National Oceanic and Atmospheric Administration, is geared toward ramp forecasts. Written in MATLAB, the Ramp Tool and Metric is a stand-alone program and is not easily modified by users. The Solar Forecast Arbiter (Hansen et al. 2019), created and maintained by a team of experts at the University of Arizona, Sandia National Laboratories, Electric Power Research Institute, and Sharply Focused, is an established Python tool for renewable energy forecast evaluation. The Solar Forecast Arbiter is open source and is equipped with a dashboard and an application programming interface that connects to its host server and its database. Users of the Solar Forecast Arbiter would upload their data to its data center and perform analysis on its server.

Compared to the other forecast evaluation tools, WE-Validate is open source, readily available, easily customizable, and computationally lightweight. WE-Validate is documented on GitHub, and we encourage and welcome contributions from the wind energy community to the tool. After installing Python and the required packages, users can download and use WE-Validate at their convenience. Users can also extend WE-Validate’s existing capabilities by adding data-processing functions, forecast evaluation metrics, ramp definitions, or visualizations to achieve their objectives. Users can write their own data-ingesting functions so that theoretically any type of data can be processed by the tool. We encourage users to write unit tests for the metrics they develop and add to WE-Validate. Moreover, users only need a local machine to execute WE-Validate, without uploading their data to a server.

# Metrics evaluation

In this section, we discuss the characteristics of different metrics for the evaluation of time-series and ramp forecasts. The wind energy community often uses single-value metrics to summarize forecast performance. Different categories of metrics are tailored for specific purposes; for example, the mean square error determines accuracy, the probability of detection targets precision (a measure of data spread), and the Peirce skill score accounts for a model’s skill relative to a reference model (Liemohn et al. 2021). Although using summary metrics is useful for making comparisons, collapsing multidimensional data into a single-number metric loses valuable information. Therefore, depending on the goal of the forecast evaluation, analysts should consider multiple metrics of different aspects for a holistic examination (Möhrlen, Zack, and Lerner 2019). For example, a suite of metrics for solar power forecasting is discussed in Zhang et al. (2015). In the following subsections, we review several commonly used metrics to evaluate time-series and ramp forecasts for wind energy applications.

## Single-value metrics for time series forecasts

This section focuses on metrics for nonprobabilistic forecasts for continuous predictands, which are appropriate for the deterministic time series evaluation of wind speed or wind power forecasts. To begin, we briefly discuss two statistical properties: robustness and resistance. A robust statistic is insensitive to assumptions made on the nature of the data, and a resistant statistic is insensitive to a small portion of outliers (Wilks 2011). For instance, on the one hand, an arithmetic mean is not robust because the mean does not adequately characterize the center of a non-Gaussian distribution and may result in misleading interpretations. An arithmetic mean is also not resistant because the mean can change drastically when a few extreme values are added to the data set, and hence the mean does not sufficiently characterize the center of the data set anymore. On the other hand, the 50th percentile of a data set, also known as the median, is robust and resistant because the median does not make any assumptions on the distribution of the data set and is not influenced by a few outliers.

The wind energy community recommends using the root-mean-square error (RMSE) to evaluate time series forecasts (Möhrlen, Zack, and Lerner 2019; Draxl et al. 2019). However, the RMSE is neither robust nor resistant because it involves the arithmetic mean. When a wind power forecast fails to predict power fluctuations for a short period, which often takes place during ramp events, the RMSE of the forecast can be overly inflated thanks to its nonresistance to outliers. Researchers have been using variations of RMSE to mitigate RMSE’s weaknesses, such as normalized RMSE (Chakraborty and Elzarka 2018) and unbiased RMSE (Entekhabi et al. 2010), but the augmentations do not fundamentally resolve its lack of statistical robustness and resistance.

Experts from other fields, such as space weather forecasting and soil sciences, have proposed metrics based on the relative magnitude between forecast and observed values, such as mean absolute percentage error and its variant like the mean arctangent absolute percentage error (Kim and Kim 2016). Median symmetric accuracy is an example of such metrics that is both robust and resistant (Morley, Brito, and Welling 2018; Liemohn et al. 2021):

However, metrics based on the ratio between forecast and observation are not ideal for wind power forecast evaluation. First, such a ratio for an observation-forecast pair of 20 MW and 10 MW and the ratio for another pair of 200 MW and 100 MW are the same. The fractional metric does not convey the magnitude of the forecast error, which can have substantial financial implications in wind energy applications. Second, when the observed power is 0 MW, the ratio generates mathematical errors from division by 0. Even though median symmetric accuracy has many valuable traits, we caution readers on metrics that use relative magnitude between forecast and observed values.

To conclude, we suggest readers use robust and resistant metrics in their analyses in addition to their established workflow using typical forecast evaluation metrics. One example that we employed in this analysis, which is robust and resistant and which preserves the magnitude of the variable, is the median absolute error. We incorporated the median absolute error and other common metrics such as the RMSE, the mean bias, and the mean absolute error in the existing version of WE-Validate.

## Metrics for ramp forecasts

This section focuses on metrics for nonprobabilistic forecasts for discrete predictands, which are appropriate for evaluating deterministic wind ramp event forecasts. Wind ramp events add power-generation variability and pose challenges to the grid, and a library of studies has been dedicated to wind ramp detection, forecasting, and evaluation metrics (Gallego-Castillo, Cuerva-Tejero, and Lopez-Garcia 2015; Cui et al. 2017; Zhang et al. 2017; Hannesdóttir and Kelly 2019; Cheneka, Watson, and Basu 2020; Dorado-Moreno et al. 2017; Ouyang et al. 2017; Sevlian and Rajagopal 2013; Ferreira et al. 2011). Researchers at the National Oceanic and Atmospheric Administration also developed a software package, the Ramp Tool and Metric, that uses a set of ramp detection definitions and sophisticated skill scores to evaluate ramp forecasts (Bianco et al. 2016). In this section, we discuss several commonly used ramp metrics that have been incorporated into WE-Validate (Section 2.0).

We use a 2x2 contingency table to evaluate deterministic wind ramp forecasts compared to observations, where the four categories are true positive (TP) or hit, false positive (FP) or false alarm, false negative (FN) or miss, and true negative (TN). Mathematical combinations of the four categories yield useful scalar attributes for ramp forecast evaluation, and we list several that are discussed in this manuscript:

which is the ratio of correct forecasts to observed ramps, and a forecast with a higher POD is more favorable;

which is the percentage of forecast ramps that are wrong, and a forecast with a lower false alarm ratio is more favorable;

which is the percentage of forecast ramps that are right, and a forecast with a higher SR is more favorable;

which is the ratio of forecast ramps to observed ramps. An unbiased forecast yields a bias of unity, a forecast that over-forecasts ramps yields a bias larger than 1, and a forecast that under-forecasts ramps yields a bias smaller than 1;

which is the ratio of correct forecasts to the total number of forecast and observed ramps. The worst possible forecast yields a CSI of 0, and the best possible forecast yields a CSI of unity. Regarding forecasting rare events when occurrences take place fewer than nonoccurrences, such as wind ramps, CSI is useful because it does not account for TN, which would be a relatively large number compared to the other three categories (Wilks 2011). The geometric relationship among POD, SR, bias, and CSI is discussed in Roebber (2009), which can be visualized as a performance diagram (Figure 16).

Note that the false alarm ratio differs from the false alarm rate (Barnes et al. 2009; Wilks 2011):

which is the ratio of false alarms to the total number of nonramp instances.

Besides the scalar attributes mentioned above, we recommend using equitable scalar skill scores to evaluate wind ramp forecasts. An equitable skill score rates random forecasts and forecasts of constant results equally, where the skill score for a useless forecast is usually defined to be zero, and a perfect forecast often yields a skill score of unity (Gandin and Murphy 1992; Wilks 2011). Equitability also implies that correct forecasts of less frequent events have more weights than correct forecasts of more common events (Wilks 2011). An example of an equitable skill score is the Peirce skill score (PSS):

where a perfect forecast yields a PSS of unity, a random forecast yields a PSS of 0, and a forecast worse than a random forecast yields a negative PSS. When ramp events are rare, a correct ramp forecast contributes more to the PSS.

In addition to the PSS, the symmetric extreme dependency score (SEDS) is another useful metric for ramp forecasts (Hogan et al. 2010; Hogan, O’Connor, and Illingworth 2009; Ferro and Stephenson 2011):

where n is the total number of deterministic ramp forecasts, and . A perfect forecast yields a SEDS of unity, a random forecast yields a SEDS of 0, and a forecast inferior to a random forecast yields a negative SEDS. When TP is 0, the resultant SEDS is undefined. The SEDS is argued as asymptotically equitable, in which it approaches equitability when the data sample size is large (Hogan et al. 2010; Hogan, O’Connor, and Illingworth 2009). The SEDS is appropriate for evaluating rare event forecasts because it ignores the potentially large contribution from TN.

In the existing version of WE-Validate, we included the 2x2 contingency table, the scalar attributes, and the two skill scores discussed above. We encourage users to expand the current library of ramp metrics in WE-Validate (Section 2.0).

# Benchmark exercise

The goal of the benchmark is to demonstrate the use of the WE-Validate tool for building on the WFIP2 V&V work. The benchmark exercise aims to demonstrate the importance of reproducible, metrics-based model assessments, which should be part of every organization’s forecast validation strategy. In that sense, this benchmark exercise provides an opportunity for us to evaluate the forecast performance of numerical weather prediction models at both intra and interorganizational levels. This exercise also serves as a platform for stakeholders to share and compare wind forecast evaluation metrics among organizations. In this study, we have further used the benchmark results to illustrate the utility of a variety of metrics. The purpose of this exercise is not to determine the most accurate forecast, but to illustrate the value of a systematic forecast evaluation framework.

Setting up a rigorous forecast evaluation procedure also aligns with the verification and validation framework proposed in the WFIP2 (Draxl et al. 2019) as well as the IEA Wind Task 36 Recommended Practice for Forecasting Solution Selection (Möhrlen, Zack, and Lerner 2019). Forecast providers in the wind energy industry as well as wind energy researchers were invited to participate in this exercise. Participants were given several months in 2021 to prepare and submit their forecasts. The authors of this manuscript from the National Renewable Energy Laboratory and the Pacific Northwest National Laboratory in the United States organized and coordinated this benchmark exercise.

After we collected data from the participants, we anonymously evaluated the submitted data and executed WE-Validate for each submission (Section 2.0). For the data analysis, we used statistics such as the RMSE and the median absolute error (Section 3.1), as well as more sophisticated skill scores for wind ramp events (Section 3.2). We also varied the configuration files depending on the submissions to test for sensitivity of the methodology (e.g., we investigated the influence of averaging frequency to resultant forecast errors [Section 5.1]). In the long run, we hope that WE-Validate will become a useful reference forecast evaluation framework for the wind energy community.

Table 1 describes the metadata of the two case studies that participants could submit data for: the WFIP2 case at the Columbia River Basin in the Pacific Northwest of the United States and the Baltic-2/FINO2 case in the North Sea in Europe.

Table 1. Summary of the two cases

|  |  |  |
| --- | --- | --- |
| Case study | WFIP2 | Baltic-2/FINO2 |
| Site description | The WFIP2 project was a meteorological measurement field campaign targeting the U.S. Pacific Northwest. The region has complex terrain and onshore wind farms. More information can be found in Olson et al. (2019), Wilczak et al. (2015), and Shaw et al. (2019). The location of the WFIP2 sodar is projected in Figure 1. | The Baltic-2 offshore wind farm is on the Germany side of the North Sea in Europe. The wind farm has 80 Siemens SWT-3.6-120 wind turbines, with a hub height of 78.25 m, rotor diameter of 120 m, and rated power of 3.6 MW. The plant capacity is 288 MW, and the wind farm has been operating since 2015.  The FINO2 research platform has been operating since 2007. The platform offers various measurements that support research on oceanography, meteorology, and ecology.  The locations of Baltic-2 and FINO2 are depicted in Figure 2. |
| Latitude and longitude (WGS84) of measurements | Sodar: 45.57451°N, 120.74734°W | FINO2 tower: 55.006928°N, 13.154189°E  Baltic-2 wind farm: 54.9733°N, 13.1778°E  FINO2 is about 4 km northwest of Baltic-2. |
| Evaluation period  (one initialization at the start of the forecast) | Start: 2016-09-23, 12 UTC  End: 2016-09-25, 12 UTC  A 48-hour forecast | Start: 2020-10-03, 23 UTC  End: 2020-10-10, 23 UTC  A 168-hour forecast |
| Validation measurement type | Temporal averages from a Vaisala Triton wind profiler | FINO2: Temporal averages from cup anemometers and wind vanes  Baltic-2: Wind-farm-average power and nacelle wind speed |
| Data frequency | Data are averaged at an interval of 10 minutes at the end of the bin (e.g., data labeled at 00:10 UTC represent averages from 00:00 to 00:10 UTC) | FINO2: Data are averaged at an interval of 10 minutes at the midpoint of the interval, which starts at 00:05 of the hour (e.g., data labeled at 00:05 represent averages from 00:00 to 00:10).  Baltic-2: Data are averaged at an interval of 15 minutes at the end of the bin (e.g., data labeled at 00:15 represent averages from 00:00 to 00:15). |
| Benchmark variables [units] available at heights | Wind speed [m s-1] and wind direction [degrees] at 40, 80, 120 m above ground level | FINO2: Wind speed [m s-1] at 62, 72, 82, and 92 m and wind direction [degrees] at 51, 71, 91 m above sea level  Baltic-2: Plant-level power [MW] and nacelle wind speed [m s-1] at 78.25 m above sea level |
| Meteorological description | Northwesterly flow and mountain waves were observed in the area. The area was overcast at times, with scattered showers in the Columbia River Basin during the first half of the forecast period. More details can be found in Draxl et al. (2021). | Based on the FINO2 tower data, southwesterly flow at hub height was observed for most of the 7-day period. The hub height temperature was never below freezing. Precipitation was recorded at 60 m on 4, 7, 8, 9, and 10 October. |
| Notes on wind farm(s) | Wind farms exist and operate in the area, but no wind power data were available. | Wind turbine availability was 100%. |

Map

Description automatically generated

Figure 1. The geographical location of the WFIP2 sodar in the Pacific Northwest region of the United States.

Map

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 2. The geographical locations of the FINO2 tower and the Baltic-2 wind farm in the North Sea of Europe. Because we could not share the specific turbine locations with the participants, we asked them to submit spatially averaged Baltic-2 forecasts in the area bounded by the black lines.

We asked the participants to submit 30-minute forecasts for both cases. For the WFIP2 case, we asked for wind velocity forecasts over 2 days; for the European case, we asked for wind velocity and plant-level power forecasts over 7 days. We asked the participants to submit forecasts aligning with the metadata of the observations in Table 1, which allowed for valid comparisons between forecasts and observations as well as comparisons among forecasts. We also asked the participants to provide metadata of their numerical models, including the resolutions of the model grid cell and the differences between the ensemble members.

We briefly summarize the submissions we received in Table 2. Participant p3 did not submit data for the WFIP2 case, and Participant p5 submitted forecasts at 60-minute intervals. For a consistent assessment among organizations, we analyzed and presented 60-minute averages for all the forecast and observed data in this study. Note that the data we gathered in this benchmark exercise are not strictly forward-looking weather forecasts because the participants could use historical reanalysis data to initialize their numerical models. The submissions of Participants p2 to p5 are referred to as ensembles because they provided more than one modeled forecast, whereas p1 only submitted a single forecast for each case study. The ensemble members use different model settings, such as various wake parameterization schemes, surface layer schemes, planetary boundary layer schemes, and vertical diffusion schemes.

Table 2. Summary of collected forecast submissions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Participant | p1 | p2 | p3 | p4 | p5 |
| Number of ensemble members | N/A (single submission) | 2 | 2 | 8 | 75 |
| Forecast output temporal resolution (minutes) | 5 | 30 | 30 | 30 | 60 |
| Number of domains | 3 | WFIP2: 4  Baltic-2/FINO2: 2 | 3 | 2 | Participants did not share this information |
| Grid resolution | 25, 5, and 1 km | WFIP2: 13, 6.5, 3.2, and 1.6 km  Baltic-2/FINO2: 13 and 6.5 km | 18, 6, and 2 km | 9 and 3 km | WFIP2: 0.15°  Baltic-2/FINO2: 0.225° |
| Number of vertical levels | 109 | 60 | 80 | 35 | 32 |

# Analysis of benchmark submissions

## Time series forecasts

The submitted forecasts were analyzed and compared to observations for both benchmark case studies. Except for p1, we calculated the ensemble means for other participants. We then treated the ensemble means of the participants, including the single-member forecasts from p1, as members of one multiorganization ensemble, which allowed us to calculate a five-organization ensemble mean. Thus, for each case study, we compared the skills of individual ensemble members, intraorganization ensemble means, and the five-organization ensemble mean.

Often, an ensemble mean yields a below-average forecast error compared to those of its members (Figure 4). As expected, the five-organization ensemble mean has weaker temporal fluctuations than the intra-organization ensemble means (Figure 3), which may have boosted its forecast performance. Across the two case studies, the five-organization ensemble mean performs better than most, and sometimes all, of the other submitted forecasts, both in terms of the RMSE and the median absolute error over the whole forecast periods (Figure 4). The superiority of the ensemble means in single-value summary metrics is even more apparent in wind power time series forecasts. The nonlinear power curve conversion results in satisfactory power forecast performance of the ensemble means compared to their individual members (Figure 4).

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Figure 3. (top) Wind speed time series of the WFIP2 case at 80 m above ground level, where the red line illustrates cup anemometer measurements, the magenta line denotes the p1 forecast, the purple, light blue, and cyan lines, respectively, indicate ensemble mean forecasts from p2, p4, and p5, and the black line is the ensemble average of the four participants. (middle) Similar to (top), but for wind speed of the FINO2 case at 82 m above sea level with submissions from five participants. (bottom) Similar to (middle), but for wind-farm power data at Baltic-2 at 78.25 m above sea level. Across the panels, the shading around the ensemble means of p4 and p5 represents the standard deviation of the ensemble members around the mean. The ensemble means of p2 and p3 are plotted as single lines and no shading is incorporated because the differences between the two ensemble members are mostly trivial.

The choice of the evaluation metrics affects the relative skills between forecasts. Compared to their ensemble members, ensemble means sometimes yield larger relative errors using median absolute error than RMSE, and this pattern is particularly apparent in the FINO2 case (Figure 4). The disparity of relative errors between the two metrics emerges from the large magnitude of errors of outliers. For RMSE, squaring the error at each time step magnifies the impacts of those outliers and creates a long tail in the squared-error distribution. Therefore, an ensemble mean would yield a lower relative RMSE than its members, whereas the same ensemble mean derives a relatively modest median absolute error.

A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot

Description automatically generated

A picture containing scatter chart

Description automatically generated

Figure 4. Swarm plots of root-mean-square error (left) and median absolute error (right) of the p4 ensemble (blue), the p5 ensemble (cyan), and the five-organization ensemble (gray) on the forecasts of WFIP2 wind speeds at 80 m above ground level (top), FINO2 wind speeds at 82 m above sea level (middle), and Baltic-2 hub height power (bottom). In each data column, each dot represents an ensemble member, and the cross indicates the respective ensemble average.

During the 7-day period at the Baltic-2 wind farm, the five-organization ensemble mean is never the worst wind power forecast among the members at any given hour, and occasionally it is the best of all (Figure 5). More than half of the time, the ensemble mean ends up in the middle of the pack, ranked 3 or 4, which fits our expectation. Along the same line, the performance of the ensemble mean is above average for more than half the time. Meanwhile, the individual forecasts of all five organizations have been the worst for some periods, and p5 is ranked first for more than a quarter of the time.

A picture containing text, transport, traffic light, vector graphics

Description automatically generated

Figure 5. The rank of the absolute error of hourly power forecast among six time series: p1, the four intraorganization ensemble means, and the five-organization ensemble mean (top left as “avg”) at the Baltic-2 wind farm. A member is ranked 1 for an hour when its power forecast error is the lowest among all members, including the ensemble mean, at that hour. Each pie chart illustrates the portion of the ranks over the 7-day period that each member holds.

Even its overall performance is above average, the error distribution of the five-organization ensemble mean does not wildly differ from those of its members. The underlying absolute error distributions of wind power forecasts are analogous to the probability density function of an exponential distribution, where the magnitude of most errors is small and close to 0 (Figure 6). To examine whether the absolute error distributions statistically differ from each other, we use the two-sample Kolmogorov-Smirnov test for each pair of the distributions. Based on the Kolmogorov-Smirnov test with an alpha of 0.05, p4’s absolute error distribution is significantly different from the other distributions except for p3’s. The absolute error distribution of the five-organization ensemble mean is also statistically different from p4’s. Therefore, even the five-organization ensemble mean often yields above-average results (Figure 5), we cannot conclude that its error distribution differs from those of all the members.

Graphical user interface

Description automatically generated

Figure 6. Histograms of absolute errors of the hourly power forecast over the 7-day period at the Baltic-2 wind farm. The five-organization ensemble mean is marked as “avg.”

To investigate why the ensemble means yield lower errors than their members, we transformed the time series of hourly power forecast errors into power spectra. Using a power spectrum, we can understand how each forecast performs during wind power fluctuations of various frequencies. The integral of the spectral components across all the frequencies corresponds to the error variance of the time series. Therefore, when the integrated power spectral density is lower than the others, the associated RMSE is also lower than the others.

The five-organization ensemble mean smooths out the extremes of its members, leading to lower power forecast errors, and a spectrum of lower integrated magnitude (Figure 7). For example, the wind power patterns fluctuating at about 2.1 hours are better captured by the five-organization ensemble mean, and hence its power errors at that frequency are lower than those of its members. Similar features also emerge at fluctuations of about 2.5 and 3.7 hours. The spectra of the p4 and p5 ensembles also display that the single-organization ensemble means have below-average integrated magnitude. The p4 ensemble mean has a smaller spectral component integral than all its individual members, and only five p5 ensemble member yields a slightly smaller spectral component integral of the same order of magnitude than the p5 ensemble mean. To summarize, an ensemble mean only needs to perform better than most its members at several frequencies to generate low forecast errors.

A picture containing shape

Description automatically generated

Chart

Description automatically generated with medium confidence

Chart

Description automatically generated with medium confidence

Figure 7. Power spectra of the hourly power forecast error of the Baltic-2 case over the 7-day period for the p4 ensemble (top), the p5 ensemble (bottom), and the five-organization ensemble (bottom). In each spectrum, the black line indicates the ensemble mean and the other colored lines represent the ensemble members.

We also investigated whether different averaging time frames of a time series would lead to different error distributions. For instance, using the same FINO2 82-m wind speed forecasts (e.g., p5 ensemble members), we performed the two-sample T-test on a pair of RMSE distributions at 60-minute and 120-minute frequencies. We found that the resultant p-value does not exceed any meaningful alpha threshold (not shown). Therefore, we could not reject the null hypothesis that the two RMSE distributions have identical averages. We performed the two-sample T-tests on other error metrics presented in this manuscript, such as the median absolute error, at different heights and we drew the same conclusions. Thus, in this work, the error distributions are independent of different averaging time scales.

## Ramp forecasts

In this study, we define a wind ramp event when the absolute change in power in a given period exceeds a threshold. For instance, the ramp definition of |50 MW| within 6 hours means that a change in power above 50 MW—either a positive or negative change—in any 6-hour interval is labeled as a ramp event, and such a definition applies to forecast and observed power time series. We implemented this simple ramp definition in WE-Validate to display the capability of the code base, and users are welcome to add their own ramp definitions. We summarize the occurrences of observed wind power ramps and no-ramps of the Baltic-2 case in Table 3.

Table 3. Observed ramp counts with different ramp definitions at the Baltic-2 wind farm during the 7-day period

|  |  |  |
| --- | --- | --- |
| Ramp definition | Observed ramp count | Observed no-ramp count |
| |50 MW| in 2 hours | 57 | 109 |
| |50 MW| in 4 hours | 80 | 84 |
| |50 MW| in 6 hours | 82 | 80 |
| |50 MW| in 8 hours | 89 | 71 |
| |100 MW| in 2 hours | 17 | 149 |
| |100 MW| in 4 hours | 31 | 133 |
| |100 MW| in 6 hours | 38 | 124 |
| |100 MW| in 8 hours | 45 | 115 |
| |100 MW| in 10 hours | 35 | 123 |
| |150 MW| in 4 hours | 9 | 155 |
| |150 MW| in 6 hours | 14 | 148 |
| |150 MW| in 8 hours | 17 | 143 |
| |150 MW| in 10 hours | 18 | 140 |
| |150 MW| in 12 hours | 21 | 135 |
| |200 MW| in 4 hours | 3 | 161 |
| |200 MW| in 6 hours | 3 | 159 |
| |200 MW| in 8 hours | 7 | 153 |
| |200 MW| in 10 hours | 9 | 149 |
| |200 MW| in 12 hours | 10 | 146 |

Given a ramp definition, a comparison of deterministic ramps between a pair of forecast and observed power time series yields a 2x2 contingency table. Each square in Figure 8 projects a contingency table of a ramp definition and a forecast-observation pair, where the two upper left triangles are correct forecasts (TP and TN), and the two lower right triangles denote incorrect forecasts (FP and FN). The numbers and the colors display the counts of the four categories in the contingency table. A skillful forecast would yield higher counts in the upper-left triangles than the lower-right triangles. Combining multiple contingency tables at once in Figure 8 enables us to contrast the skills of various forecasts under different ramp definitions.

In our Baltic-2 case study, the multiorganization ensemble mean records fewer correct power-ramp forecasts than its members. The submissions from three of the five organizations (p1, p2, and p3) are individually better at correctly forecasting ramps, or TP, above 50 MW over 4, 6, and 8 hours than the five-organization ensemble mean (Figure 8). For ramps above 50 MW over 2 and 4 hours, the ensemble mean has above-average correct nonramp forecasts, or TN. For ramps of longer durations, 50 MW over 6 and 8 hours, the ensemble mean leads FN ramp forecasts, where the ensemble mean fails to predict ramps. For false alarms, or FP, the performance of the ensemble mean is above average across ramp definitions.

Background pattern

Description automatically generated

Figure 8. Illustration of the 2x2 contingency table on the power ramp forecast at the Baltic-2 wind farm over the 7-day period. Each row represents a ramp event definition, from changing over |50 MW| within 2 hours to changing over |50 MW| within 8 hours, and the ramp definitions include both up ramps and down ramps. Each column is an ensemble mean, except for p1, which submitted single-member forecasts. The four triangles in each square characterize the 2x2 contingency table, as percentages of instances among the four parameters. The percentages are annotated in the triangles. The sum of the four triangles in each square is 100%.

Combining the ramp detection of individual forecasts often leads to better ramp forecasts than detecting ramps with the ensemble mean forecast time series. The ensemble mean forecast smooths out temporal fluctuations, and such removal of peaks and troughs cripples the ensemble mean in adequately predicting ramp events. To further examine this phenomenon, we implement a voting scheme between ensemble members to detect ramps. In the following paragraphs and figures in this section, for a five-member ensemble, the “40% vote” scheme tags a period as a ramp forecast when two of the five ensemble members forecast a ramp under a ramp definition. For instance, only p3 and p5 forecast ramp from 6 UTC to 12 UTC on 5 October, so the 40% voting scheme labels that period as a forecast ramp (the gray area in Figure 9). In the same example, the 20% voting scheme also tags the period as a ramp forecast because at least one of the five voting members indicates ramp. For all the individual ramp forecasts as well as the various voting schemes, we use the same observed ramps to compute the 2x2 contingency table.

Chart, histogram

Description automatically generated

Figure 9. Illustration of determining wind power ramps with voting scheme using a ramp definition of |50 MW| within 6 hours for the Baltic-2 case. In this example, two out of five voting members indicate ramps between 6 UTC and 12 UTC on 5 October, so the 40% voting scheme labels ramp in this period (gray). Similarly, five out of five voting members indicate ramps between 17 UTC and 23 UTC on 5 October, and three out of five voting members indicate ramps between 4 UTC and 10 UTC on 6 October, so the respective 100% (yellow) and 60% (green) voting schemes label ramps for the two periods.

The voting schemes vary in strengths and weaknesses. As expected, the 20% voting scheme is the most sensitive in detecting ramps, which scores the highest in TP as well as FP ramp forecasts among all forecasts listed in Figures 10 and 11. In contrast, the 80% voting scheme is stringent in indicating ramps, resulting in substantially higher FN and lower FP than the others (Figures 11 and 12). The 60% voting scheme has a relatively high threshold-to-ramp detection, and even it yields more TP and fewer FN ramp forecasts for ramps above 50 MW at 6 and 8 hours than the ensemble mean. Like the 60% voting scheme, the 40% voting scheme largely increases TP and decreases FN compared to the ensemble mean, but it also increases FP.

In the Baltic-2 case, more episodes of observed power fluctuations are labeled as ramp events (sum of TP and FN) when the length of a ramp detection period increases. In the model forecasts, TP and FP ramp forecasts for ramps above 50 and 100 MW also share such a pattern, whereas TN ramp forecasts monotonically decrease with increasing ramp-detection duration for all ramp definitions (Figure 13).

Calendar

Description automatically generated

Figure 10. The counts of true positive wind power ramp forecasts over the 7-day period at the Baltic-2 wind farm. Each row represents a ramp event definition, and all ramp definitions include both up ramps and down ramps. From left to right, the columns illustrate different voting schemes (XX% vote), the five-organization ensemble mean (all 5p mean), and the forecasts from each organization (pX or pX mean).

A screenshot of a computer

Description automatically generated

Figure 11. Similar to Figure 10, but for false positive forecasts.

Calendar

Description automatically generated

Figure 12. Similar to Figure 10, but for false negative forecasts.

Calendar

Description automatically generated

Figure 13. Similar to Figure 10, but for true negative forecasts.

To summarize the outcomes of the 2x2 contingency table, we employ the PSS, which is an equitable metric. The five-organization ensemble mean is not much better than random forecasts in ramp forecasting, except for ramps above 100 and 150 MW over 10 hours (Figure 14). Except for p5, the forecasts of the other four participants yield more positive PSS than the five-organization ensemble mean, especially for stronger ramps above 150 and 200 MW. Mathematically, the relative magnitude of the higher POD of the four individual forecasts exceeds the influence of the higher false alarm rate, which leads to higher PSS than the multiorganization ensemble mean (Figure 16 as an example).

The voting schemes score higher using the PSS than the five-organization ensemble mean, especially for the ramps above 200 MW. The 40% voting scheme has a higher or same PSS compared to the ensemble mean in all but one ramp definition (50 MW over 4 hours). Even the stricter 60% voting scheme yields higher PSS than the ensemble mean in most ramp definitions. The 20% voting scheme has very high PSS in detecting ramps over 200 MW, while its exposes a weakness of the PSS. The PSS is inflated when FN is close to 0, the false alarm rate is close to 0, and the POD is large, or its FN is 0 for ramps above 200 MW over 6 and 12 hours. Similar features also exist in p4 ensemble mean.

The SEDS is a valuable complement to the PSS. The SEDS does not account for any TN, which is suitable when ramp events are rarer than nonramp periods. Using SEDS, the relative performance of the ensemble mean improves from its PSS results (Figure 15). The edge of the four participant forecasts over their ensemble mean also shrinks, and they are still more skillful in detecting the ramps above 200 MW. Note that an undefined SEDS means the forecast does not yield any TP forecasts under a given ramp definition.

The more sensitive voting schemes also lose some superiority over the five-organization ensemble mean using SEDS. The 20% and 40% voting schemes yield higher SEDS than the ensemble mean in detecting strong ramps above 200 MW, and they lose advantages in most of the other ramp definitions. The 60% voting scheme has a higher or same SEDS than the ensemble mean except for one ramp definition (50 MW over 2 hours). The 80% voting scheme is too inflexible that it forecasts zero TP in one more ramp definition than the ensemble mean (Figure 10).

Table, calendar

Description automatically generated

Figure 14. Similar to Figure 10, but for the Peirce skill score using the 2x2 contingency table projected in Figures 10, 11, 12, and 13.

Calendar, treemap chart

Description automatically generated

Figure 15. Similar to Figure 10, but for the symmetric extreme dependency score using the 2x2 contingency table projected in Figures 10, 11, 12, and 13. A blank cell means that the SEDS is undefined, wherein the forecast does not yield any TP forecasts under a given ramp definition.

We further use performance diagrams to contrast the skills of multiple deterministic wind power ramp forecasts on the same chart. A performance diagram uses the geometric relationship among four scalar attributes of the 2x2 contingency table—the POD, the SR, the bias, and the CSI—and projects the information onto one diagram (Roebber 2009). A performance diagram assists us in comparing ramp forecasting abilities of the ensemble mean, individual ensemble members, and different voting schemes (Figure 16).

Changing a voting scheme to another involves modifying the sensitivity of ramp detection, and such a switch follows a pattern on a performance diagram. Theoretically, increasing the sensitivity of ramp detection (reducing the percentage of votes required to indicate ramps) increases the instances of TP or FP or both. Therefore, mathematically, raising this sensitivity increases bias and POD. Because both TP and FP are in the denominator of SR and CSI, adjusting this sensitivity has an undetermined impact on these two parameters. As a result, a voting scheme that is sensitive to ramp detection (e.g., 20% vote) always has a higher POD and a higher bias than those insensitive to ramp detection, and the relative magnitude of their SR and CSI depends on the forecasts and the ramp definition.

Choosing a voting scheme can more correctly and effectively forecast ramp events than using the ensemble mean. In the five-organization example for ramps above 50 MW over 6 hours (Figure 16), all the ensemble members and voting schemes have a higher POD, a higher SR, and a higher CSI than the ensemble mean. Among all voting schemes, the 60% voting scheme achieves a satisfactory balance among POD, SR, bias, and CSI. In the p5-ensemble example (Figure 16), the 50% voting scheme appears close to the center of the ensemble members on the performance diagram, whereas the ensemble mean tends to underforecast ramps and have fewer TP.

A picture containing chart

Description automatically generated

A screenshot of a web

Description automatically generated with low confidence

Figure 16. Performance diagrams of the Baltic-2 power-ramp forecasts for the 7-day period under the ramp definition of changing over |50 MW| within 6 hours. The x-axis is success ratio, the y-axis is probability of detection, the dashed diagonal lines extending from the origin represent bias values, and the blue contours indicate a range of critical success index values. A perfect forecast would land on the top-right corner of the diagram. The top diagram depicts the ensemble means of each participant as purple dots, the results of voting schemes as green triangles, and the five-organization ensemble mean as a black cross. The bottom diagram illustrates the p5 ensemble members as gray dots, the results of voting schemes among the p5 ensemble members as triangles, and the p5 ensemble mean as a black cross.

# Discussions

In this effort, we designed and hosted a benchmark exercise, and we gathered and analyzed wind power forecasts. Through this benchmark, we demonstrated the importance of a rigorous validation process and the need to consider various metrics in determining the value of a forecast. We evaluated the wind speed and wind power forecasts in two geological and climatic regions, as well as in onshore and offshore environments. We also identified and documented forecast evaluation methodologies with an open-source code base, WE-Validate. The significance of this benchmark exercise also lies in data sharing as well as knowledge sharing among collaborators. Through this exercise and WE-Validate, we aim to improve the value of wind energy forecasts to the wind energy industry.

WE-Validate provides a platform and a set of forecast evaluation steps for analysts to use and refer to. When forecast providers modify their operational workflow, such as changing the input data, advancing data assimilation techniques, updating model physics, and shifting forecasting horizons, they can test the forecast improvements via a systematic framework. Users can also use WE-Validate to select the ideal forecast providers (Möhrlen and Zack 2019; Möhrlen, Zack, and Lerner 2019). The analysts can also equip the code base with their own modules and functionalities to fit their purposes. With this validation tool, the wind energy community can fairly assess forecasts of different models and organizations in a coherent manner.

We discuss the importance of using statistically robust and resistant metrics to evaluate time-series forecasts as well as equitable metrics to evaluate deterministic ramp forecasts. When evaluating forecasts, we advise accounting for multiple metrics for a comprehensive analysis. If outliers exist in the data set, the choice of metrics can affect the relative errors between forecasts and thus the deduced conclusions. For instance, relative to its members, an ensemble mean can display superb skills with RMSE while yielding a modest median absolute error. Furthermore, analysts should spend time understanding the characteristics of the metrics they choose; for instance, the PSS accounts for TN while the SEDS does not. Depending on the goals, analysts should consider various ramp definitions in the analysis. Sometimes strong ramps have larger financial implications than weak ramps, and the same applies to down ramps compared to up ramps.

We explore the strengths and weakness of ensemble means in this study. An ensemble mean shaves off the extreme forecasts of its members, and its smooth pattern acts as a double-edged sword. In time-series forecast evaluation using single-value metrics such as RMSE and median absolute error, ensemble means often have acceptable errors and do not yield the most errors among all ensemble members at any time steps. The wisdom of the crowd prevails in time-series forecasts because an ensemble mean squashes the differences and underscores the common features between ensemble members.

However, ensemble means underperform in ramp forecasts. In deterministic ramp forecast evaluations using the 2x2 contingency table, ensemble means become less skillful than many of their ensemble members, assuming the members have adequate ramp forecasting skills. Ramp forecasts of a voting system among individual ensemble members on indicating ramps often yield more correct ramp predictions than an ensemble mean, where the latter has a higher chance of missing forecast ramps. A sensitive voting scheme also leads to a larger number of forecast ramps. With a voting scheme, a trade-off usually exists between larger FP and more TP with fewer FN. A sensitive voting scheme can be useful when missing extreme ramp events (FN) brings costly consequences. In contrast to time series forecasting, the wisdom of the crowd carries a different meaning in ramp forecasting, where member voting at each time step becomes advantageous in ramp detection. Ultimately, the skill of an ensemble mean as well as a voting scheme is dictated by the skill of the ensemble members. Skillful ensemble members yield a skillful ensemble mean.

# Conclusions

Through this study, we established and showcased a code base, WE-Validate, to evaluate multiple wind forecasts in a consistent fashion. WE-Validate, written in Python, is open-source, modularized, and extensible by users. We select two case studies in a benchmark exercise to exhibit the systematic forecast evaluation procedure layout in WE-Validate. Participants from industry and academia engaged in the benchmark exercise and contributed to its success, and we analyzed their data submissions via WE-Validate in this work.

We discuss the importance of employing statistically robust and resistant metrics as well as equitable skill scores. We analyzed the collected data with a median absolute error, an example of a robust and resistant metric, and the Peirce skill score, an example of an equitable skill score. We also recommended using the symmetric extreme dependency score, an asymptotically equitable metric, to evaluate forecasts of rare events such as wind ramps.

We further investigated the performance of the ensemble mean. We found that the ensemble mean has adequate skill in time-series forecasting and underperforms in ramp forecasting compared to its ensemble members. The spectral analysis suggests that the ensemble mean performs better than most of its members at several frequencies, thus generating low time-series forecast errors for the whole forecast period. In ramp forecasting, the 2x2 contingency table reveals that ensemble means tend to miss predicting ramps. We develop an arrangement where individual ensemble members vote to detect ramps at a given time step, and such a voting scheme gains skills in ramp forecasting, especially for ramps of large magnitude. The downside of a sensitive voting scheme is risking more false alarms, while its likely benefit is resulting in more hits with fewer misses.

Looking forward, we hope that the wind energy industry acknowledges the advantages of robust and resistant forecast validation. The next phase of this study includes adding capabilities to WE-Validate to evaluate probabilistic forecasts and to quantify forecast uncertainty with a systematic procedure. We welcome community contribution to WE-Validate to refine the wind forecast validation process.

# References

Barnes, Lindsey R., David M. Schultz, Eve C. Gruntfest, Mary H. Hayden, and Charles C. Benight. 2009. “CORRIGENDUM: False Alarm Rate or False Alarm Ratio?” *Weather and Forecasting* 24 (5): 1452–54. https://doi.org/10.1175/2009WAF2222300.1.

Bessa, Ricardo J., Corinna Möhrlen, Vanessa Fundel, Malte Siefert, Jethro Browell, Sebastian Haglund El Gaidi, Bri Mathias Hodge, Umit Cali, and George Kariniotakis. 2017. “Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry.” *Energies 2017, Vol. 10, Page 1402* 10 (9): 1402. https://doi.org/10.3390/EN10091402.

Bianco, Laura, Irina V. Djalalova, James M. Wilczak, Joel Cline, Stan Calvert, Elena Konopleva-Akish, Cathy Finley, and Jeffrey Freedman. 2016. “A Wind Energy Ramp Tool and Metric for Measuring the Skill of Numerical Weather Prediction Models.” *Weather and Forecasting* 31 (4): 1137–56. https://doi.org/10.1175/WAF-D-15-0144.1.

Chakraborty, Debaditya, and Hazem Elzarka. 2018. “Performance Testing of Energy Models: Are We Using the Right Statistical Metrics?” *Journal of Building Performance Simulation* 11 (4): 433–48. https://doi.org/10.1080/19401493.2017.1387607.

Cheneka, Bedassa R., Simon J. Watson, and Sukanta Basu. 2020. “A Simple Methodology to Detect and Quantify Wind Power Ramps.” *Wind Energy Science* 5 (4): 1731–41. https://doi.org/10.5194/WES-5-1731-2020.

Craig, Anna, Mike Optis, Michael Jason Fields, and Patrick Moriarty. 2018. “Uncertainty Quantification in the Analyses of Operational Wind Power Plant Performance.” *Journal of Physics: Conference Series* 1037 (5): 052021. https://doi.org/10.1088/1742-6596/1037/5/052021.

Cui, Mingjian, Jie Zhang, Cong Feng, Anthony R. Florita, Yuanzhang Sun, and Bri Mathias Hodge. 2017. “Characterizing and Analyzing Ramping Events in Wind Power, Solar Power, Load, and Netload.” *Renewable Energy* 111 (October): 227–44. https://doi.org/10.1016/J.RENENE.2017.04.005.

Dorado-Moreno, M., L. Cornejo-Bueno, P. A. Gutiérrez, L. Prieto, C. Hervás-Martínez, and S. Salcedo-Sanz. 2017. “Robust Estimation of Wind Power Ramp Events with Reservoir Computing.” *Renewable Energy* 111 (October): 428–37. https://doi.org/10.1016/J.RENENE.2017.04.016.

Draxl, Caroline, 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. https://doi.org/NREL/TP-5000-72553.

Draxl, Caroline, Rochelle P. Worsnop, Geng Xia, Yelena Pichugina, Duli Chand, Julie K. Lundquist, Justin Sharp, Garrett Wedam, James M. Wilczak, and Larry K. Berg. 2021. “Mountain Waves Can Impact Wind Power Generation.” *Wind Energy Science* 6 (1): 45–60. https://doi.org/10.5194/WES-6-45-2021.

Entekhabi, Dara, Rolf H. Reichle, Randal D. Koster, and Wade T. Crow. 2010. “Performance Metrics for Soil Moisture Retrievals and Application Requirements.” *Journal of Hydrometeorology* 11 (3): 832–40. https://doi.org/10.1175/2010JHM1223.1.

Ferreira, C., J. Gama, L. Matias, A. Botterud, and J. Wang. 2011. “A Survey on Wind Power Ramp Forecasting.” Argonne, IL (United States). https://doi.org/10.2172/1008309.

Ferro, Christopher A. T., and David B. Stephenson. 2011. “Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events.” *Weather and Forecasting* 26 (5): 699–713. https://doi.org/10.1175/WAF-D-10-05030.1.

Gallego-Castillo, Cristobal, Alvaro Cuerva-Tejero, and Oscar Lopez-Garcia. 2015. “A Review on the Recent History of Wind Power Ramp Forecasting.” *Renewable and Sustainable Energy Reviews* 52 (December): 1148–57. https://doi.org/10.1016/J.RSER.2015.07.154.

Gandin, Lev S., and Allan H. Murphy. 1992. “Equitable Skill Scores for Categorical Forecasts.” *Monthly Weather Review* 120 (2): 361–70. https://doi.org/10.1175/1520-0493(1992)120.

Giebel, Gregor, William Shaw, Helmut Frank, Caroline Draxl, Pierre Pinson, George Kariniotakis, and Corinna Möhrlen. 2018. “Task Proposal Extension of Task 36 Forecasting for Wind Energy, Phase II.”

Hannesdóttir, Ásta, and Mark Kelly. 2019. “Detection and Characterization of Extreme Wind Speed Ramps.” *Wind Energy Science* 4 (3): 385–96. https://doi.org/10.5194/WES-4-385-2019.

Hansen, Clifford W., William F. Holmgren, Aidan Tuohy, Justin Sharp, Antonio T. Lorenzo, Leland J. Boeman, and Anastasios Golnas. 2019. “The Solar Forecast Arbiter: An Open Source Evaluation Framework for Solar Forecasting.” *Conference Record of the IEEE Photovoltaic Specialists Conference*, June, 2452–57. https://doi.org/10.1109/PVSC40753.2019.8980713.

Hogan, Robin J., Christopher A. T. Ferro, Ian T. Jolliffe, and David B. Stephenson. 2010. “Equitability Revisited: Why the ‘Equitable Threat Score’ Is Not Equitable.” *Weather and Forecasting* 25 (2): 710–26. https://doi.org/10.1175/2009WAF2222350.1.

Hogan, Robin J., Ewan J. O’Connor, and Anthony J. Illingworth. 2009. “Verification of Cloud-Fraction Forecasts.” *Quarterly Journal of the Royal Meteorological Society* 135 (643): 1494–1511. https://doi.org/10.1002/QJ.481.

Kim, Sungil, and Heeyoung Kim. 2016. “A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts.” *International Journal of Forecasting* 32 (3): 669–79. https://doi.org/10.1016/J.IJFORECAST.2015.12.003.

Lee, Joseph C.Y., and M. Jason Fields. 2021. “An Overview of Wind-Energy-Production Prediction Bias, Losses, and Uncertainties.” *Wind Energy Science* 6 (2): 311–65. https://doi.org/10.5194/WES-6-311-2021.

Liemohn, Michael W., Alexander D. Shane, Abigail R. Azari, Alicia K. Petersen, Brian M. Swiger, and Agnit Mukhopadhyay. 2021. “RMSE Is Not Enough: Guidelines to Robust Data-Model Comparisons for Magnetospheric Physics.” *Journal of Atmospheric and Solar-Terrestrial Physics* 218 (July): 105624. https://doi.org/10.1016/J.JASTP.2021.105624.

Möhrlen, Corinna, and John Zack. 2019. “Expert Group Report on Recommended Practices for Selecting Renewable Power Forecasting Solutions - Part 1: Forecast Solution Selection Process.”

Möhrlen, Corinna, John Zack, and Jeffrey Lerner. 2019. “Expert Group Report on Recommended Practices For Selecting Renewable Power Forecasting Solutions - Part 2: Designing and Executing Forecasting Benchmarks and Trials.”

Morley, S. K., T. V. Brito, and D. T. Welling. 2018. “Measures of Model Performance Based On the Log Accuracy Ratio.” *Space Weather* 16 (1): 69–88. https://doi.org/10.1002/2017SW001669.

Olson, Joseph B., Jaymes S. Kenyon, Irina Djalalova, Laura Bianco, David D. Turner, Yelena Pichugina, Aditya Choukulkar, et al. 2019. “Improving Wind Energy Forecasting through Numerical Weather Prediction Model Development.” *Bulletin of the American Meteorological Society* 100 (11): 2201–20. https://doi.org/10.1175/BAMS-D-18-0040.1.

Ouyang, Tinghui, Xiaoming Zha, Liang Qin, Yi Xiong, and Heming Huang. 2017. “Model of Selecting Prediction Window in Ramps Forecasting.” *Renewable Energy* 108 (August): 98–107. https://doi.org/10.1016/J.RENENE.2017.02.035.

Reback, Jeff, Wes McKinney, jbrockmendel, Joris van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, et al. 2020. “Pandas-Dev/Pandas: Pandas 1.0.3 (v1.0.3).” Zenodo. https://doi.org/10.5281/zenodo.3715232.

Roebber, Paul J. 2009. “Visualizing Multiple Measures of Forecast Quality.” *Weather and Forecasting* 24 (2): 601–8. https://doi.org/10.1175/2008WAF2222159.1.

Sevlian, Raffi, and Ram Rajagopal. 2013. “Detection and Statistics of Wind Power Ramps.” *IEEE Transactions on Power Systems* 28 (4): 3610–20. https://doi.org/10.1109/TPWRS.2013.2266378.

Shaw, William J., Larry K. Berg, Joel Cline, Caroline Draxl, Irina Djalalova, Eric P. Grimit, Julie K. Lundquist, et al. 2019. “The Second Wind Forecast Improvement Project (WFIP2): General Overview.” *Bulletin of the American Meteorological Society* 100 (9): 1687–99. https://doi.org/10.1175/BAMS-D-18-0036.1.

Wilczak, James, Cathy Finley, Jeff Freedman, Joel Cline, Laura Bianco, Joseph Olson, Irina Djalalova, et al. 2015. “The Wind Forecast Improvement Project (WFIP): A Public–Private Partnership Addressing Wind Energy Forecast Needs.” *Bulletin of the American Meteorological Society* 96 (10): 1699–1718. https://doi.org/10.1175/BAMS-D-14-00107.1.

Wilczak, James M., Mark Stoelinga, Larry K. Berg, Justin Sharp, Caroline Draxl, Katherine McCaffrey, Robert M. Banta, et al. 2019. “The Second Wind Forecast Improvement Project (WFIP2): Observational Field Campaign.” *Bulletin of the American Meteorological Society* 100 (9): 1701–23. https://doi.org/10.1175/BAMS-D-18-0035.1.

Wilks, Daniel S. 2011. *Statistical Methods in the Atmospheric Sciences*. Amsterdam, Netherlands: Academic Press.

Zhang, Jie, Mingjian Cui, Bri Mathias Hodge, Anthony Florita, and Jeffrey Freedman. 2017. “Ramp Forecasting Performance from Improved Short-Term Wind Power Forecasting over Multiple Spatial and Temporal Scales.” *Energy* 122 (March): 528–41. https://doi.org/10.1016/J.ENERGY.2017.01.104.

Zhang, Jie, Anthony Florita, Bri Mathias Hodge, Siyuan Lu, Hendrik F. Hamann, Venkat Banunarayanan, and Anna M. Brockway. 2015. “A Suite of Metrics for Assessing the Performance of Solar Power Forecasting.” *Solar Energy* 111 (January): 157–75. https://doi.org/10.1016/J.SOLENER.2014.10.016.

Pacific Northwest   
National Laboratory

902 Battelle Boulevard

P.O. Box 999

Richland, WA 99354

1-888-375-PNNL (7665)

[***www.pnnl.gov***](http://www.pnnl.gov)