Renewable Energy xxx (xxxx) 1–18

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

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

Joseph C.Y. Lee [a], Caroline Draxl [b, *], Larry K. Berg [a]

[a] Pacific Northwest National Laboratory, Richland, WA, 99354, USA
[b] National Renewable Energy Laboratory, Golden, CO, 80401, USA

ARTICLE INFO

Article history:
Received 25 February 2022
Received in revised form 1 June 2022
Accepted 25 September 2022

Keywords:
Wind energy
Benchmark exercise
Forecast evaluation
Open-source tool
Ramp forecasting

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.

1. 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 toforecast uncertainty communications [1]. Varying methodologies chosen by different organizations also create uncertainties in predicting wind energy production [2,3]. Creating a benchmark to evaluate forecast performance is also costly for forecast users and providers [4]. Therefore, to minimize the misalignment of expectations and requirements in wind energy forecasts among stakeholders, a comprehensive and objective process of selecting forecast providers has been proposed [4,5].

Adhering to the recommended practice guidelines, 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 (WE standing for wind energy), 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 effort [6] of the Second Wind Forecast Improvement Project (WFIP2) [7–9] 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 Energy 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 energy industry [10]. 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

A2e, Atmosphere to Electrons; API, Application programming interface; CSI, Critical success index; DOE, U.S. Department of Energy; FN, False negative; FP, False positive; IEA, International Energy Agency; MET, Model Evaluation Tools; POD, Probability of detection; PSS, Peirce skill score; RMSE, Root-mean-square error; SEDS, Symmetric extreme dependency score; SR, Success ratio; TN, True negative; TP, True positive; WFIP2, Second Wind Forecast Improvement Project; WE, Wind energy

* Corresponding author.
E-mail address: Caroline.Draxl@nrel.gov (C. Draxl).

https://doi.org/10.1016/j.renene.2022.09.111
0960-1481/© 20XX

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 land-based wind farms. More information can be found in Refs. [8,9,38]. The location of the WFIP2 sodar is projected in Fig. 1.	The Baltic-2 offshore wind farm is on the Germany side of the Baltic Sea in Europe. The wind farm has 80 S 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 Fig. 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, 1200 UTC End: 2016-09-25, 1200 UTC A 48-h forecast	Start: 2020-10-03, 2300 UTC End: 2020-10-10, 2300 UTC A 168-h 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 min 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 min 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 min 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 Ref. [39].	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 2020. Multiple frontal systems passed through the Baltic Sea region in the 7-day period.
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%.

Wind Task 36, this study benefitted from the invaluable input from task members, and WE-Validate can be directly linked to real-world applications. In particular, the Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions [4,5] were consulted during the design and conception phase of the benchmark exercise and code development.

2. 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 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 useable, open source, publicly available, modularized, and extensible by users. Detailed instructions for users can be found on its GitHub page, https://github.com/a2edap/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 [11], 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; and 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 2 × 2 contingency table. When the variable of interest is wind speed and the height level matches the specified hub height, the code derives power using a power curve, which can be selected by the user. Note that the visualizations displayed in this analysis (Section 5) are not part of the standard visual outputs in the current version of WE-Validate. The diagrams presented in this work are created for this study, based on the numerical outputs from executing WE-Validate with forecast submissions of the benchmark exercise. We also made the Python code of these charts and analyses publicly available on GitHub.

Users can 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-min resolution and another forecast at 30-min resolution, the analyst can execute WE-Validate at a 30-min resolution for both forecasts and compare their performance in a compatible manner.

We note that other useful validation tools also exist or are under development. WindSider (windsider.io) is tailored for the wind resource assessment process and uses the data structure of xarray, another Python package. As of this writing, WindSider is under development, led by experts from 3E and the Technical University of Denmark. The

Table 2
Summary of collected forecast submissions.

Participant	p1	p2	p3	p4	p5	p6
Number of ensemble members	N/A (single submission)	2	2	8	75	N/A (single submission)
Forecast output temporal resolution (min)	5	30	30	30	60	30
Number of domains	3	WFIP2: 4 Baltic-2/FINO2: 2	3	2	3	1 for all models used
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°	WFIP2: 0.25° FINO2: 0.15° Baltic-2: 0.15°, 0.15°, 0.156° by 0.234°, and 0.25°
Number of vertical levels	109	60	80	35	32	WFIP2: 137 FINO2: 38 Baltic-2: 38, 38, 70, and 137
Type of forecast	Hindcast	Pre-operational	Forecast	Forecast	Forecast	Forecast

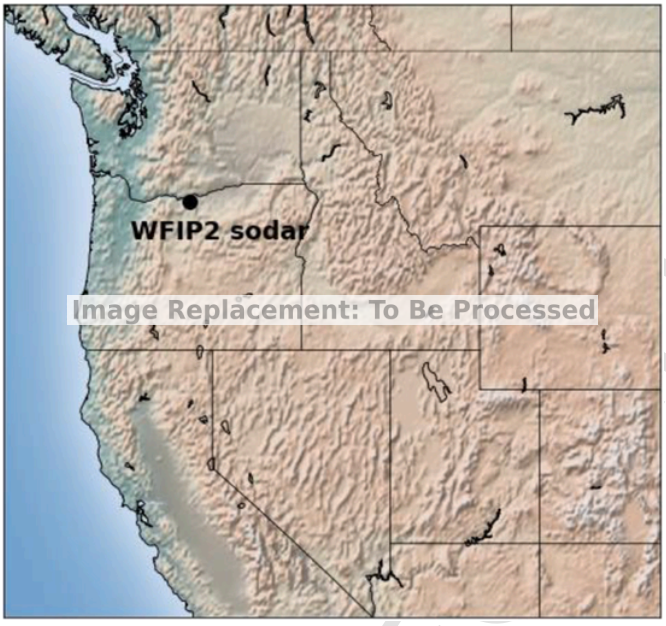

Fig. 1. The location of the WFIP2 sodar on the aerial map of the northwestern United States.

Ramp Tool and Metric [12], 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 cannot not be easily modified by users. The Solar Forecast Arbiter [13], created and maintained by a team of experts at the University of Arizona, Sandia National Laboratories, Electric Power Research Institute, Inc., and Sharply Focused LLC, 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 (API) 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. METplus [14] is a sophisticated model verification framework for numerical weather prediction. The tool is developed based on the Model Evaluation Tools (MET) and has a suite of Python wrappers for statistical and graphical analyses. METplus is developed and supported by the Developmental Testbed Center, which involves experts from the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the U.S. Air Force.

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 to the tool from the wind energy community. 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.

3. Metrics and 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 perfor-

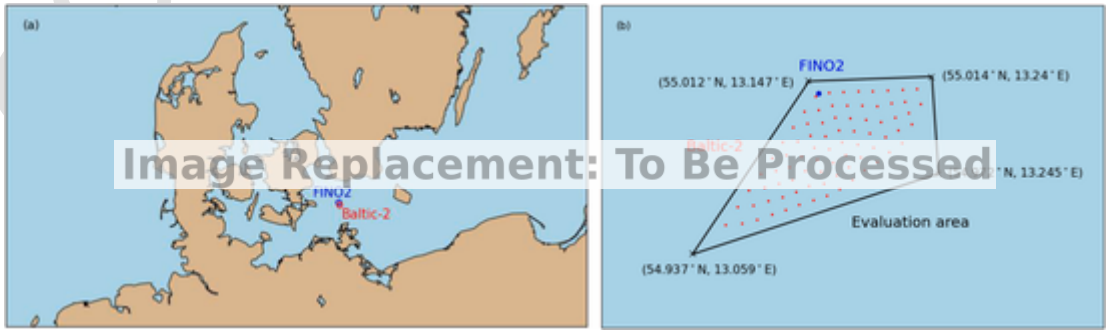

Fig. 2. The locations of the FINO2 tower (blue) and the Baltic-2 wind farm (red) on the aerial map of the Baltic Sea in Europe. Because we could not share the specific turbine locations with the participants, we asked them to submit spatially averaged Baltic-2 forecasts within the area bounded by the black-color four-sided polygon.

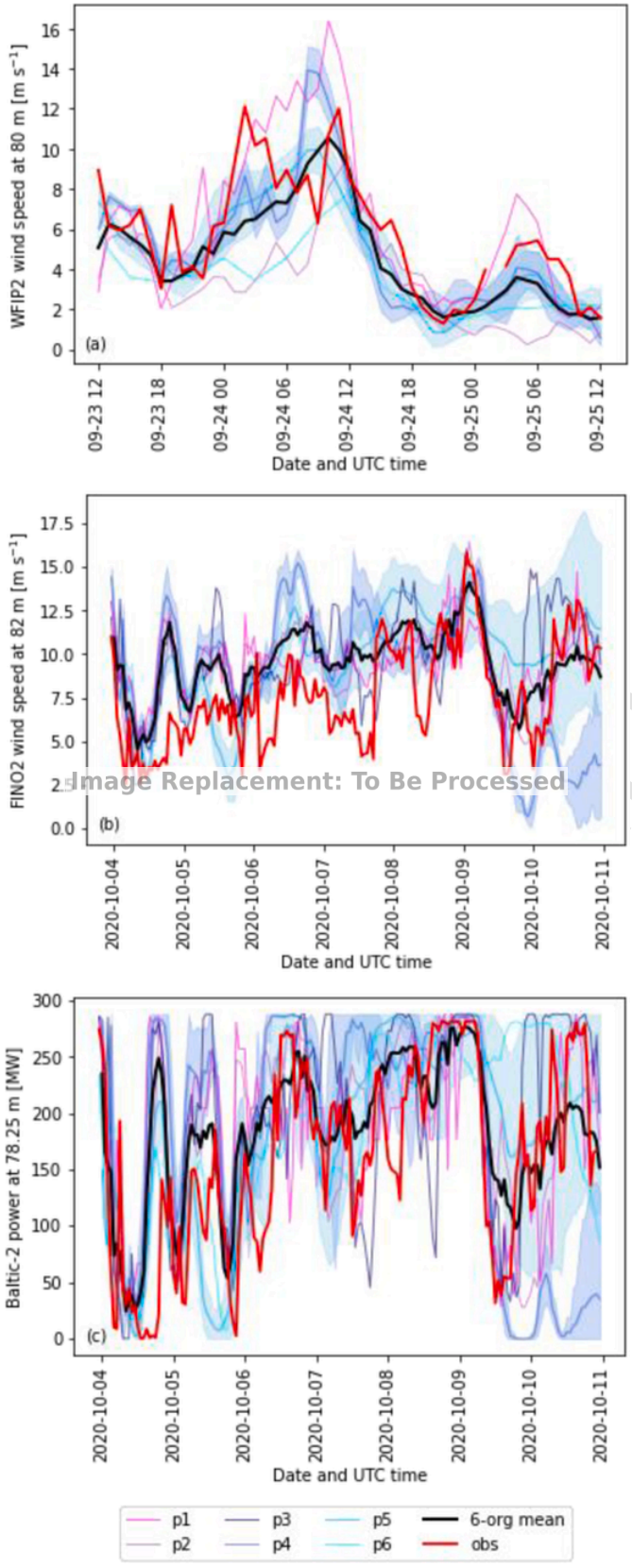

Fig. 3. (a) Wind speed time series of the WFIP2 case at 80 m above ground level, where the red line illustrates cup anemometer measurements, the magenta and cyan lines denote the p1 and p6 forecasts, the purple, blue, and light blue lines, respectively, indicate ensemble mean forecasts from p2, p4, and p5, and the black line is the ensemble average of the five participants. (b) Similar to (a), but for wind speed of the FINO2 case at 82 m above sea level with submissions from six participants. (c) Similar to (b), but for wind farm power data at Baltic-2 at 78.25 m above sea level. The p3 ensemble mean is plotted in (b) and (c) as a navy line. 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 in each submission are mostly trivial.

mance. 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 [15]. 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 [5]. For example, suites of metrics for wind and solar power forecasting are respectively discussed in Refs. [16,17]. In the following subsections, we review several commonly used metrics to evaluate time series and ramp forecasts for wind energy applications.

3.1. 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 [18]. 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.

In the wind energy community, using the root-mean-square error (RMSE) has been a common practice to evaluate time series forecasts [6,16]. 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 [19] and unbiased RMSE [20], 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 [21]. Median symmetric accuracy is an example of such metrics that is both robust and resistant [15,22]:

$$Median\ symmetric\ accuracy = 100 \times \exp\left(median\left(\left|ln\frac{forecast_i}{observation_i}\right|\right) - 1\right)$$

Nevertheless, 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

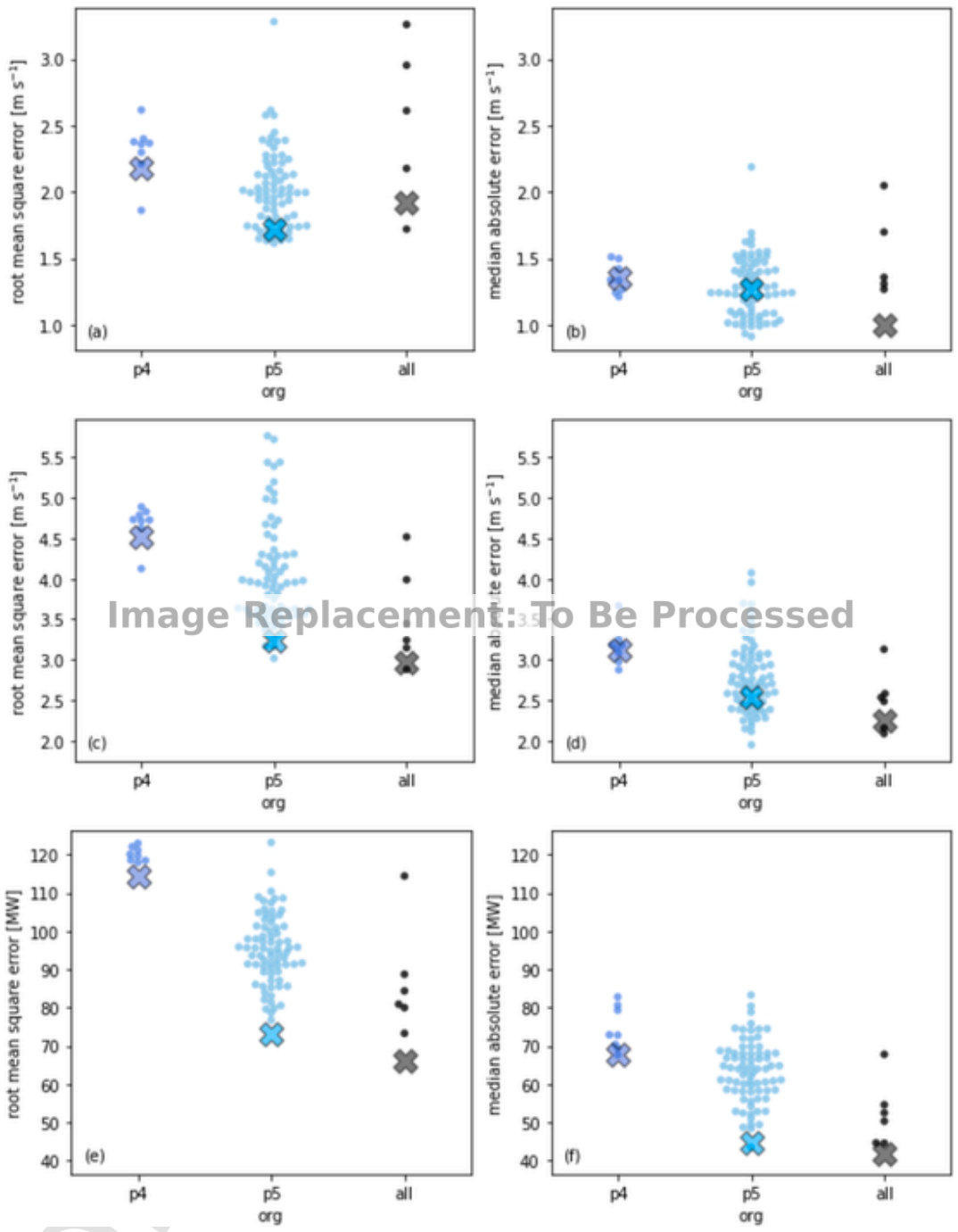

Fig. 4. Swarm plots of root-mean-square error (left column: a, c, and e) and median absolute error (right column: b, d, f) of the p4 ensemble (blue), the p5 ensemble (light blue), and the six-organization ensemble (gray) on the forecasts of WFIP2 wind speeds at 80 m above ground level (a and b), FINO2 wind speeds at 82 m above sea level (c and d), and Baltic-2 hub height power (e and f). In each data column, each dot represents an ensemble member, and the cross indicates the respective ensemble average.

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 understand the characteristics of different metrics and use robust and resistant metrics in their analyses in addition to their established workflow or commonly used metrics. One example that we employed in this analysis, which is robust and resistant and 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.

3.2. *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 many studies have been dedicated to wind ramp detection, forecasting, and evaluation metrics

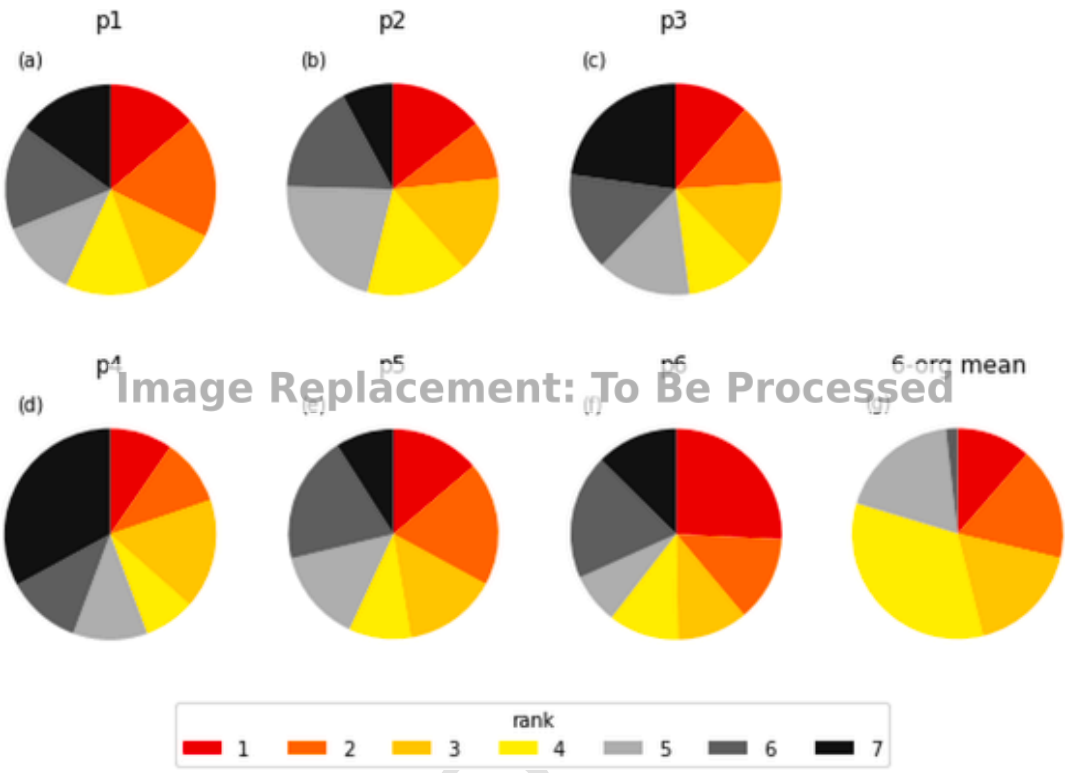

Fig. 5. The rank of the absolute error of hourly power forecast among seven time series: (a) p1; (b) p2 mean; (c) p3 mean; (d) p4 mean; (e) p5 mean; (f): p6; and (g) the six-organization ensemble mean at the Baltic-2 wind farm. A forecast is ranked 1 for an hour when its power forecast error is the lowest among all seven forecasts at that hour. Each pie chart illustrates the portion of the ranks over the 7-day period that each forecast holds.

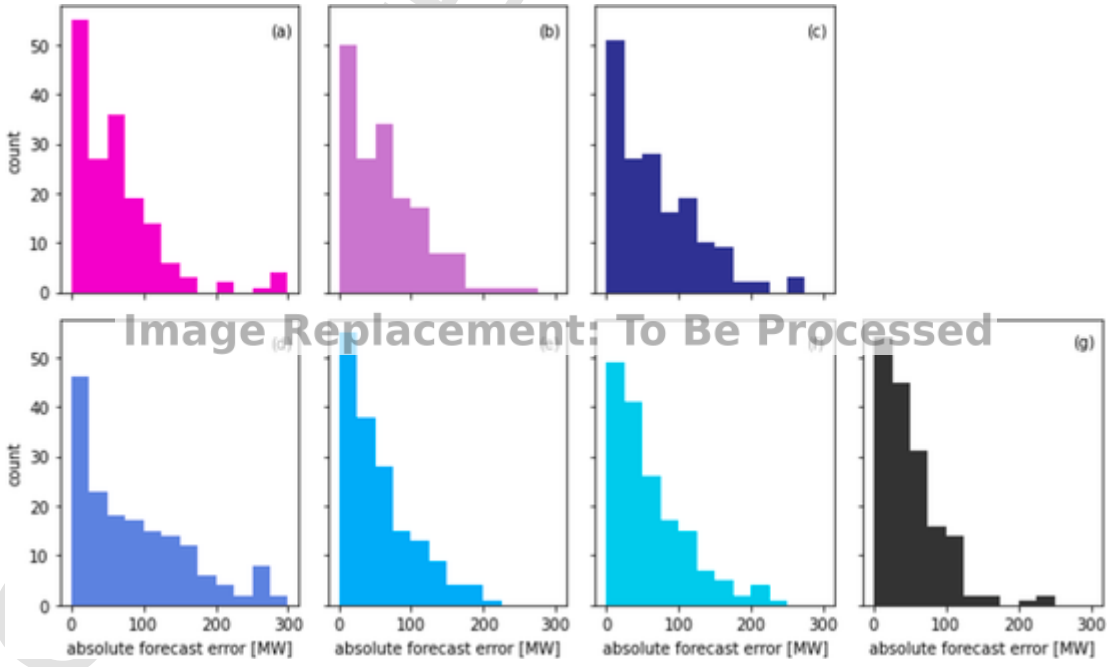

Fig. 6. Histograms of absolute errors of the hourly power forecast over the 7-day period at the Baltic-2 wind farm: (a) p1; (b) p2 mean; (c) p3 mean; (d) p4 mean; (e) p5 mean; (f) p6; and (g) the six-organization mean.

[23–31]. 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 [12]. In this section, we discuss several commonly used ramp metrics that have been incorporated into WE-Validate (Section 2).

We use a 2 × 2 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:

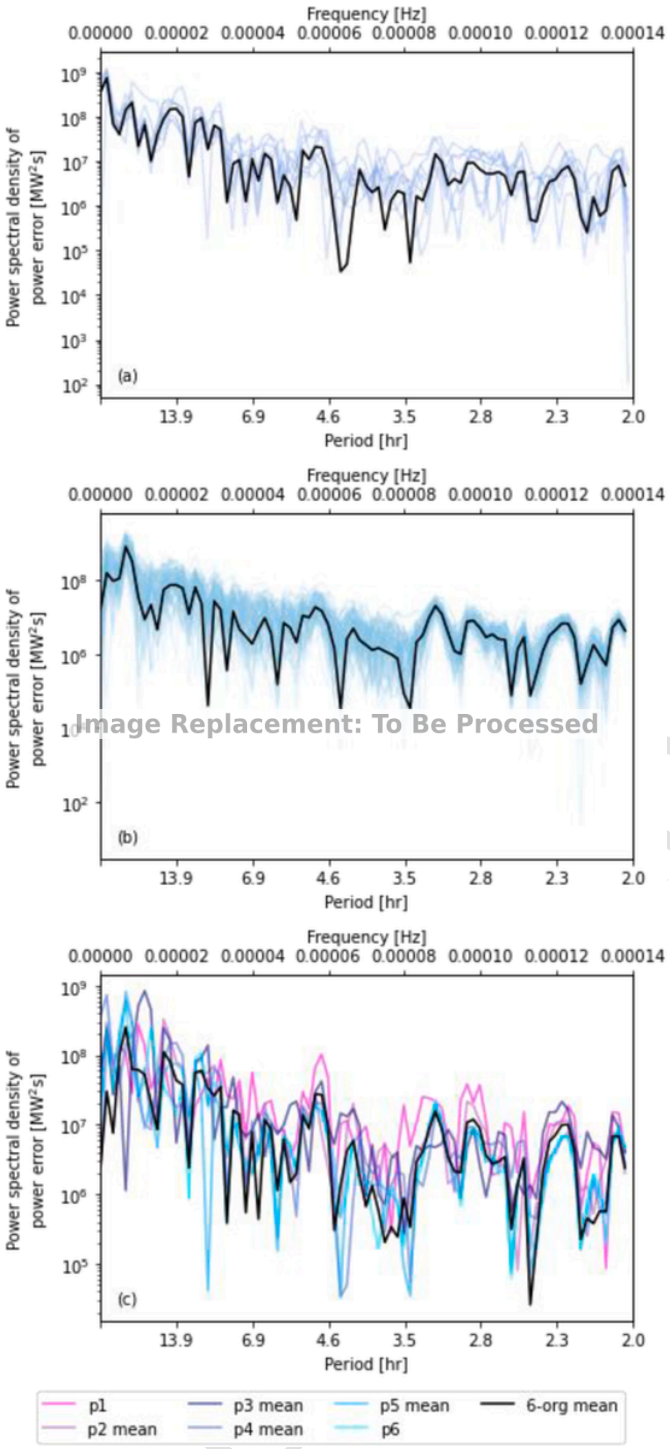

Fig. 7. Power spectra of the hourly power forecast error of the Baltic-2 case over the 7-day period for (a) the p4 ensemble, (b) the p5 ensemble, and (c) the six-organization ensemble. In each spectrum, the black line indicates the ensemble mean, and the other colored lines represent the ensemble members.

Table 3
Observed ramp counts with different ramp definitions at the Baltic-2 wind farm during the 7-day period using 60-min data.

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

$$Probability\ of\ detection\ (POD)\ or\ hit\ rate = \frac{TP}{TP+FN} \#(2)$$

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

$$False\ alarm\ ratio = \frac{FP}{TP+FP} \#(3)$$

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

$$Success\ ratio\ (SR)\ or\ forecast\ accuracy = \frac{TP}{TP+FP} = 1 - False\ alarm\ ratio \#(4)$$

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

$$Bias\ or\ frequency\ bias\ score = \frac{TP+FP}{TP+FN} = \frac{POD}{SR} \#(5)$$

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;

$$Critical\ success\ index\ (CSI)\ or\ threat\ score = \frac{TP}{TP+FP+FN} = \left(\frac{1}{SR} + \frac{1}{POD} - 1\right)^{-1} \#(6)$$

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 are 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 [18]. The geometric relationship among POD, SR, bias, and CSI is discussed in Roebber (2009), which can be visualized as a performance diagram.

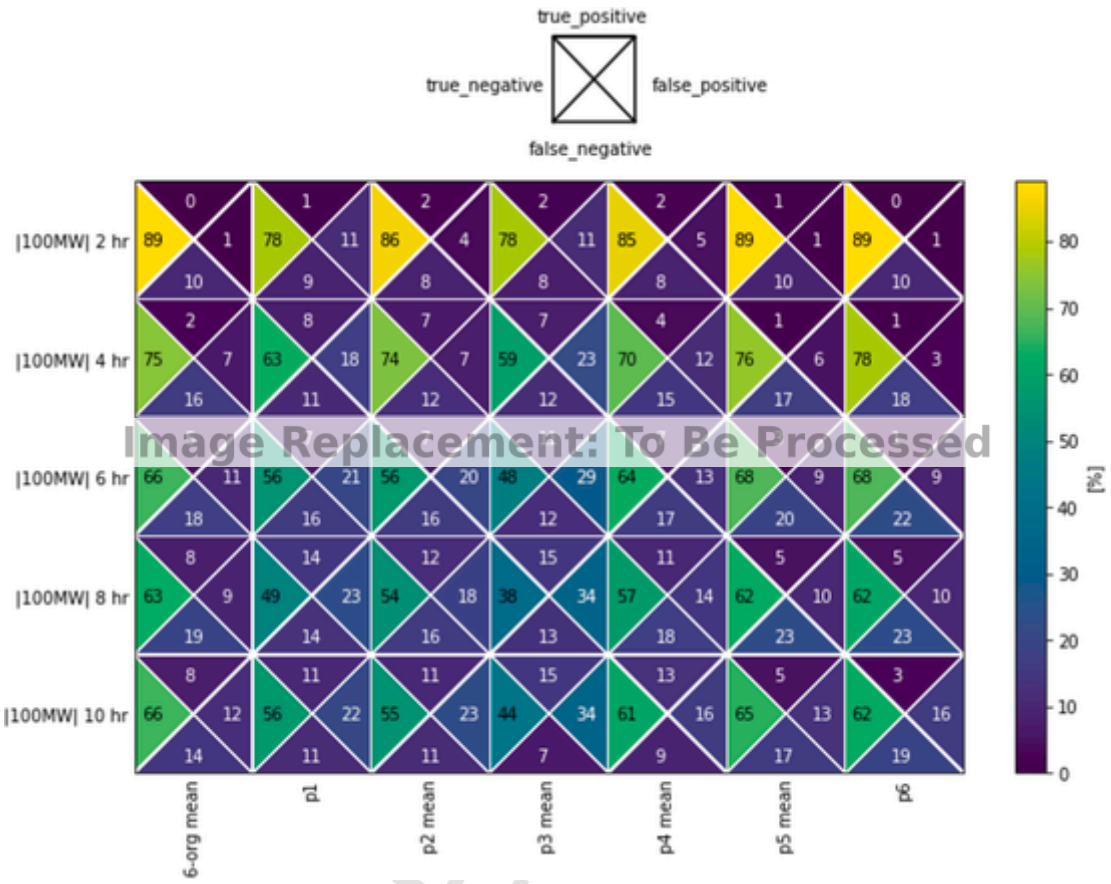

Fig. 8. Illustration of the 2 × 2 contingency table on the power ramp forecast at the Baltic-2 wind farm over the 7-day period. Each row represents a ramp event definition, from changing over |100 MW| within 2 h to changing over |100 MW| within 10 h, and the ramp definitions include both up ramps and down ramps. Each column is an ensemble mean, except for p1 and p6, which submitted single-member forecasts. The four triangles in each square characterize the 2 × 2 contingency table as percentages of instances among the four parameters in the contingency table. The percentages are annotated in the triangles. The sum of the four triangles in each square is 100%.

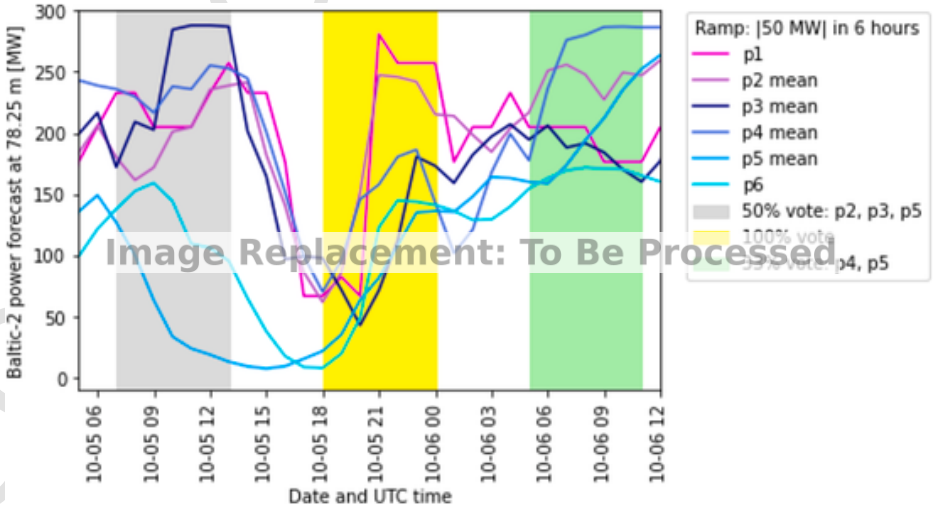

Fig. 9. Illustration of determining wind power ramps with voting scheme using a ramp definition of |50 MW| within 6 h for the Baltic-2 case in October 2020. In this example, three out of six voting members indicate ramps between 0700 UTC and 1300 UTC on 5 October, so the 50% voting scheme labels ramp in this period (gray rectangle). Similarly, six out of six voting members indicate ramps between 1800 UTC on 5 October and 0000 UTC on 6 October, and two out of six voting members indicate ramps between 0500 UTC and 1100 UTC on 6 October, so the respective 100% (yellow rectangle) and 33% (green rectangle) voting schemes label ramps for the two periods.

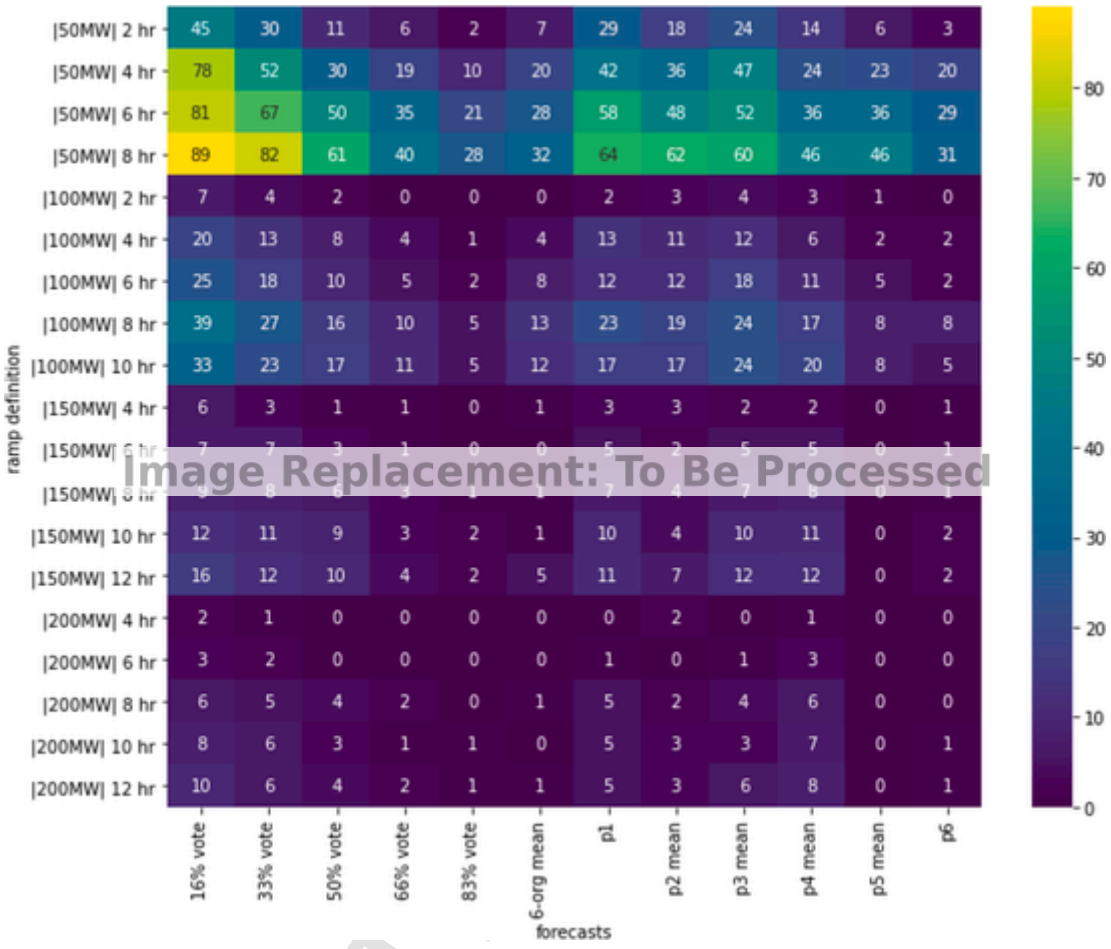

Fig. 10. The counts of true positive wind power ramp forecasts over the 7-day period at the Baltic-2 wind farm. Each row represents a ramp event definition, and all ramp definitions include both up ramps and down ramps. From left to right, the columns illustrate different voting schemes (% vote), the six-organization ensemble mean (6-org mean), and the forecasts from each organization (pX or pX mean, where X is the participant number).

Note that the false alarm ratio differs from the false alarm rate [18,33]:

$$False\ alarm\ rate = \frac{FP}{TN + FP} \#(7)$$

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 0, and a perfect forecast often yields a skill score of unity [18,34]. Equitability also implies that correct forecasts of less frequent events have more weight than correct forecasts of more-common events [18]. An example of an equitable skill score is the Peirce skill score (PSS):

$$\begin{aligned} PSS &= POD - False\ alarm\ rate \\ &= \frac{(TP \times TN) - (FP \times FN)}{(TP + FN) \times (FP + TN)} \#(8) \end{aligned}$$

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 [35–37]:

$$SEDS = \frac{\ln\left(\frac{TP+FP}{n}\right) + \ln\left(\frac{TP+FN}{n}\right)}{\ln\left(\frac{TP}{n}\right)} - 1 \#(9)$$

where n is the total number of deterministic ramp forecasts, and $n = TP + FP + FN + TN$. 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 [35,36]. 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 2 × 2 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).

4. Benchmark exercise

The goal of the benchmark is to demonstrate the use of the WE-Validate tool building on the WFIP2 model verification and validation 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

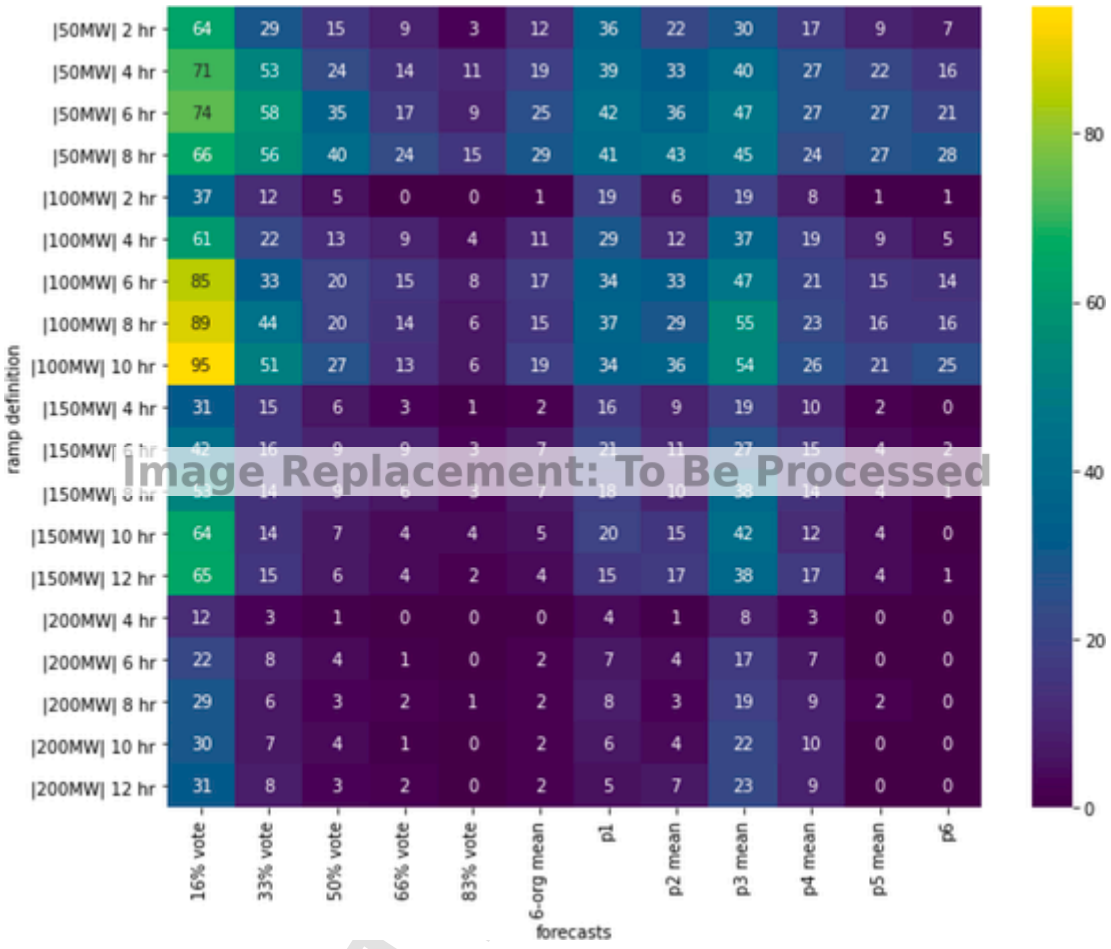

Fig. 11. Similar to Fig. 10, but for false positive forecasts.

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 [6] as well as the IEA Wind Task 36 Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions [5]. Through this work, we implemented the metrics discussed in Section 3 into WE-Validate (Section 2) to showcase the importance of the recommended framework.

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. For the data analysis, we used and contrasted commonly 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 submission, to test for sensitivity of the methodology. For instance, we investigated the influence of averaging frequency to resultant forecast errors (Section 5.1). In the long run, we hope that WE-Validate will be developed into 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.

We asked the participants to submit 30-min 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 invited 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, in the case of ensemble forecast, 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-min intervals. For a consistent assessment among organizations, we analyzed 60-min 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 from Participants p2 through 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. For example, the submission from p5 is considered a classic ensemble prediction system, in which the members differ in condensation and advection pa-

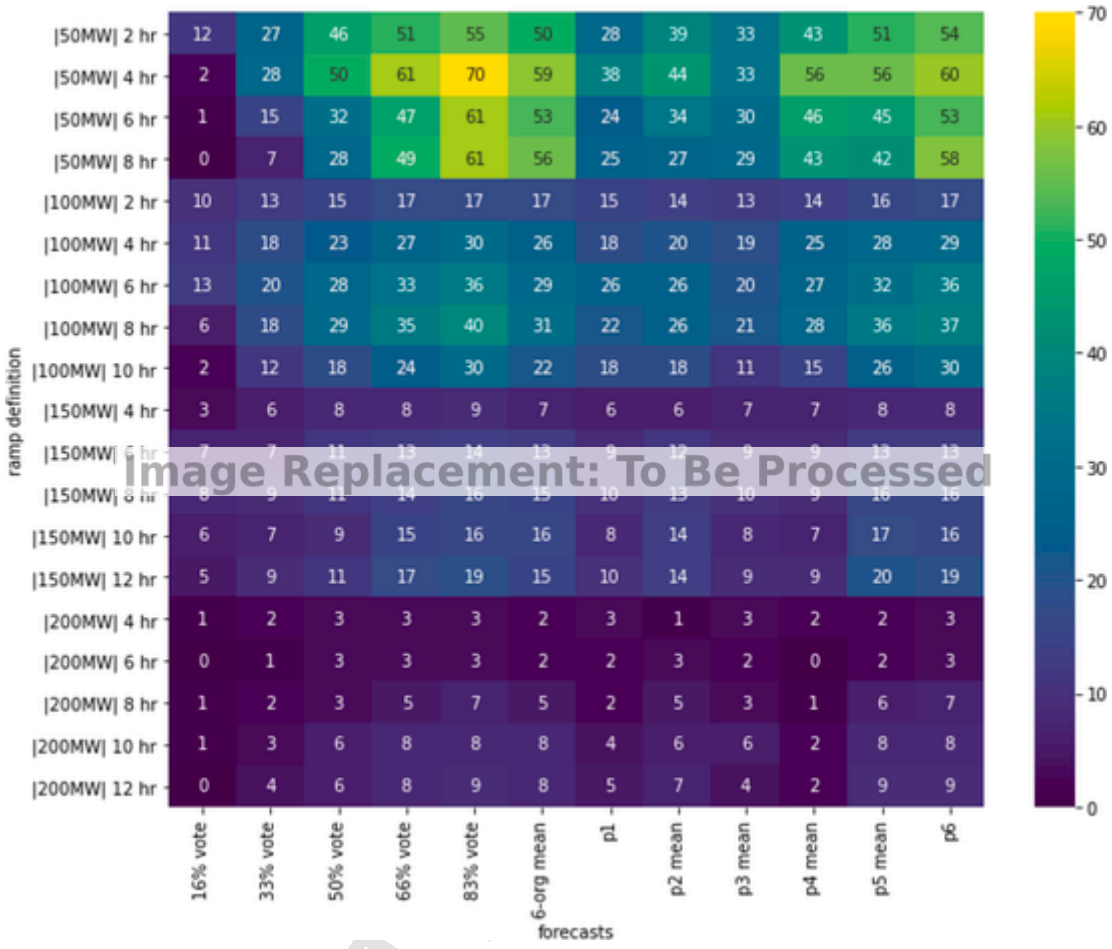

Fig. 12. Similar to Fig. 10, but for false negative forecasts.

rameterization schemes. Moreover, if a participant submitted forecasts from operation model runs, their type of forecast is a true forecast. Alternatively, p1 used reanalysis data in their simulations, and the type of forecast is a hindcast. The forecast submission from p2 involved experimental changes made to their operational model, which were then formally adopted after the submission, and thus is labeled "pre-operational" in Table 2. The Baltic-2 power forecast submission from p6 involves four different numerical weather prediction models, and their differences are listed in Table 2.

5. Analysis of benchmark submissions

5.1. Time series forecasts

In this section, we exhibit the differences among the submitted forecast data sets after we processed them with the systematic evaluation framework via WE-Validate. The forecasts were analyzed and compared to observations for both benchmark case studies (Fig. 3). Except for p1 and p6, we calculated the ensemble means for the participants. We then treated the ensemble means of the participants, including the single-member forecasts from p1 and p6, as members of one multiorganization ensemble, which allowed us to calculate a six-organization ensemble mean. Thus, for each case study, we compared the skills of individual ensemble members, intra-organization ensemble means, and the six-organization ensemble mean.

Often, an ensemble mean yields a below-average forecast error compared to those of its members (Fig. 4). As expected, the six-organization ensemble mean has weaker temporal fluctuations than the intra-organization ensemble means (Fig. 3), which may have boosted its forecast performance. Across the two case studies, the six-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 period (Fig. 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 (Fig. 4).

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 compared to using RMSE, and this pattern is particularly apparent in the FINO2 case (Fig. 4c and d). 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, because of the lack of statistical robustness and resistance of RMSE. In our case studies, given some members in an ensemble yield substantially larger errors than others, the associated ensemble mean could yield a lower relative RMSE than many of its members, whereas the same ensemble mean derives a relatively modest median absolute error.

During the 7-day period at the Baltic-2 wind farm, the six-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 (Fig. 5). More than three quarters of the time, the skill of the ensemble mean is at least average, ranked 4 or lower, which fits our expectation (Fig. 5g). Along the same line, the performance of the ensemble mean is above average for about half the time. Meanwhile, the individ-

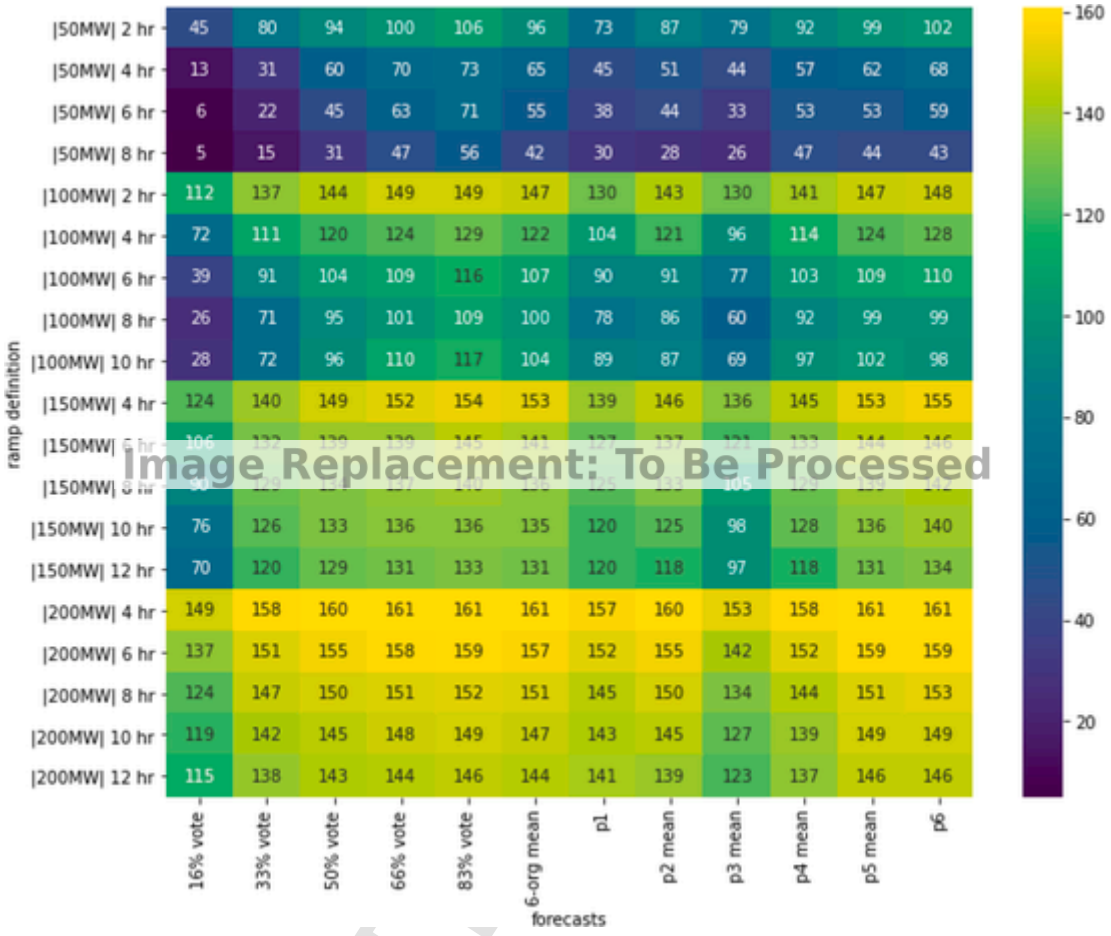

Fig. 13. Similar to Fig. 10, but for true negative forecasts.

ual forecasts of all six organizations have been the worst for some periods. Individual forecasts are often skillful as well, for instance, p6 is ranked first for about a quarter of the time (Fig. 5f). Overall, the multi-organization ensemble mean is rated above average for over 75% of the time stamps, signifying the wisdom of the crowd in time series forecasting.

Even though its overall performance is above average, the error distribution of the six-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 skewed towards 0 (Fig. 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 six-organization ensemble mean is also statistically different from p4's. Therefore, even though the six-organization ensemble mean often yields above-average forecasts (Fig. 5g), we cannot conclude that its error distribution differs from those of all the members.

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 six-organization ensemble mean smooths out the extremes of its members, leading to lower power forecast errors, and a spectrum of lower integrated magnitude (Fig. 7). For example, the wind power patterns fluctuating at about 2.5 h are better captured by the six-organization ensemble mean, and hence, its power errors at that frequency are lower than those of its members (Fig. 7c). Similar features also emerge at fluctuations of about 3.7 and 6.7 h. 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 (Fig. 7a), and only five p5 ensemble members yield slightly smaller spectral component integrals, which are of the same order of magnitude, than the p5 ensemble mean (Fig. 7b). To summarize, an ensemble mean only needs to perform better than most of its members at several frequencies to generate low forecast errors.

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-min and 120-min 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.

Fig. 14. Similar to Fig. 10, but for the Peirce skill score using the 2 × 2 contingency table projected in Figs. 10–13.

5.2. Ramp forecasts

In this section, we classify ramp events on different forecast data sets using the same set of ramp definitions through WE-Validate, and hence, we analyze the ramp forecasts in a consistent fashion. 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 h means that a change—either a positive or negative change—in power above 50 MW in any 6-h 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 using 60-min data in Table 3. When we increase the temporal resolution to 15 and 30 min and count the ramp occurrences, the resulting observed ramps and no-ramps increase, yet the ratios between them remain similar to those of the 60-min data (not shown).

5.2.1. 2 × 2 contingency table

Given a ramp definition, a comparison of deterministic ramps between a pair of forecast and observed power time series yields a 2 × 2 contingency table. Each square in Fig. 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 Fig. 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 six-organization ensemble mean tends to identify ramp events less frequently than the individual members and thus yields below-average TPs and FPs (Fig. 8). Accordingly, the ensemble generates above-average TNs and FNs. The relative magnitude among the four categories in the contingency table has implications in calculating ramp forecasting skill scores, and they are discussed later in this section.

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 attribute, we implement a voting scheme between ensemble members to detect ramps, which leads to superior ramp forecast performance to a simple ensemble mean.

5.2.2. Voting schemes

In the following paragraphs and figures in this section, for a six-member ensemble, the “50% vote” scheme tags a period as a ramp forecast when three of the six ensemble members forecast a ramp under a ramp definition. For instance, p2, p3, and p5 forecast ramp from 0700 UTC to 1300 UTC on 5 October 2020, so the 50% voting scheme labels that period as a forecast ramp (the gray area in Fig. 9). In the same example, the 33% voting scheme also tags the period as a ramp forecast because at least two of the six voting members indicates ramp. For all

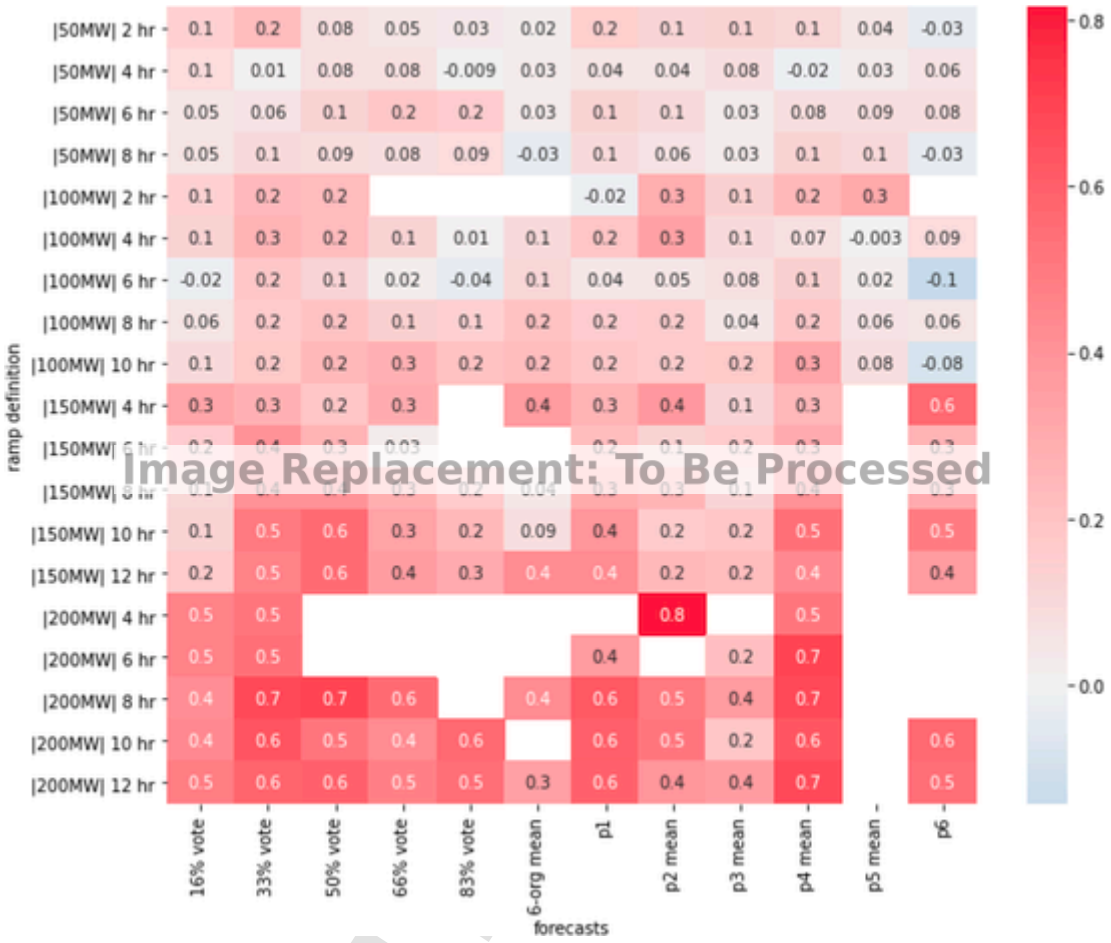

Fig. 15. Similar to Fig. 10, but for the symmetric extreme dependency score using the 2 × 2 contingency table projected in Figs. 10–13. A blank cell means that the SEDS is undefined, meaning the forecast does not yield any TP forecasts under a given ramp definition.

the individual ramp forecasts as well as the various voting schemes, we use the same observed ramps to compute the 2 × 2 contingency table.

The voting schemes vary in strengths and weaknesses. As expected, the 16% 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 Figs. 10 and 11. In contrast, the 83% voting scheme is stringent in indicating ramps, resulting in substantially higher FN and lower FP than the others (Figs. 11 and 12). The 66% 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 h than the ensemble mean. Like the 66% voting scheme, the 50% 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 (Fig. 13).

To summarize the outcomes of the 2 × 2 contingency table, we employ the PSS, which is an equitable metric. The six-organization ensemble mean generates near-zero PSSs in nearly all ramp definitions, which means it is not much more skillful than random forecasts (Fig. 14). The forecasts from p1, p2, p3, and p4 yield more positive PSSs than the six-organization ensemble mean, especially for stronger ramps above 150 and 200 MW. Mathematically, the relative magnitude of the higher POD of the four forecasts exceed the influence of the higher false alarm rate, which leads to a higher PSS than the multiorganization ensemble mean (Fig. 16 as an example). The p5 ensemble mean consists of 75 individual members and ravels their discrepancies, hence intrinsically indicating few ramps. Similarly, the p6 submission also considers four different numerical models that could also explain its above-average FN ramp forecasts.

The voting schemes score higher PSSs than the six-organization ensemble mean, especially for ramps above 200 MW (Fig. 14). The 50% voting scheme has a higher or same PSS compared to the ensemble mean in all but three ramp definitions for strong ramps in short time frames (150 MW over 4 h, 200 MW over 4 and 6 h). The stricter 66% and 83% voting schemes yields similar PSSs to the ensemble mean in most ramp definitions. Meanwhile, the 16% voting scheme has very high PSSs when detecting ramps over 200 MW. However, this pattern 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. For the 16% voting scheme, its FN counts are either 0 or 1 for ramps above 200 MW (Fig. 12), therefore its PSSs become remarkably large. Similar phenomena also exist for the 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 (Fig. 15). The edge of the participants' individual forecasts over their ensemble mean also shrinks in moderate ramps of 100 and 150 MW, and they are still more skillful in detecting the large ramps above 200 MW. Note that an undefined SEDS represents that the forecast does not yield any TP forecasts under a given ramp definition.

Using SEDS instead of PSS, the more sensitive voting schemes also lose a certain level of superiority over the six-organization ensemble mean. Because of its large FP forecasts, the SEDSs for the 16% voting

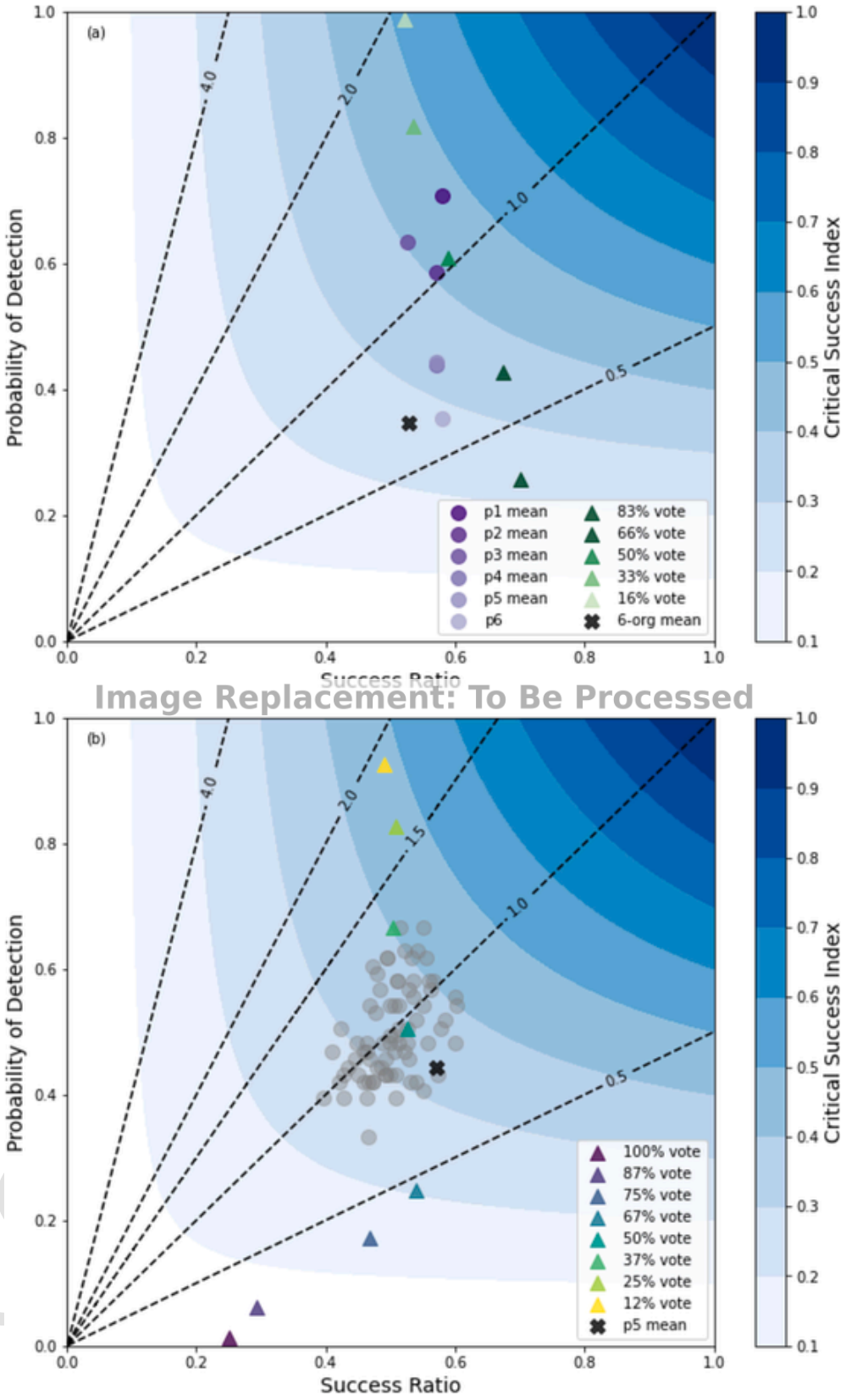

Fig. 16. Performance diagrams of the Baltic-2 power ramp forecasts for the 7-day period under the ramp definition of changing over |50 MW| within 6 h. The x-axis is the success ratio, the y-axis is the probability of detection, the dashed diagonal lines extending from the origin represent bias values, 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. (a) The ensemble means of each participant

◀ are purple dots, the results of voting schemes are green triangles, and the six-organization ensemble mean is a black cross. (b) The p5 ensemble members are gray dots, the results of voting schemes among the p5 ensemble members are triangles, and the p5 ensemble mean is a black cross.

scheme are comparable to those for the ensemble mean. The 33% voting scheme yields substantially higher SEDSs than the ensemble mean in detecting strong ramps above 200 MW, and its advantages over the ensemble mean become questionable for most of the other ramp definitions. The 50% voting scheme has a SEDS higher than or equal to the ensemble mean except for one ramp definition (150 MW over 4 h). The stringent 66% and 80% voting schemes are often inflexible, forecasting 0 TP in many ramp definitions, like the ensemble mean does (Fig. 10).

5.2.3. Performance diagram

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 2 × 2 contingency table—the POD, the SR, the bias, and the CSI—and projects the information onto one diagram [32]. A performance diagram assists us in comparing ramp forecasting abilities of the ensemble mean, individual ensemble members, and different voting schemes (Fig. 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., 16% 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 six-organization example for ramps above 50 MW over 6 h (Fig. 16a), all the ensemble members and most voting schemes have a higher POD, a higher SR, and a higher CSI than the ensemble mean. Among all voting schemes, the 50% voting scheme achieves a satisfactory balance among POD, SR, bias, and CSI. In the p5 ensemble example (Fig. 16b), the 50% voting scheme appears close to the center of the ensemble members on the performance diagram, whereas the ensemble mean tends to under forecast ramps and have fewer TPs.

6. Discussion

In this effort, we designed and hosted a benchmark exercise, and we gathered and analyzed wind power forecasts. Through this benchmark, we emphasized the importance of a rigorous validation process and the need to consider various metrics in determining the value of a forecast or forecast ensemble. We evaluated the wind speed and wind power forecasts in two geological and climatic regions, as well as in land-based and offshore environments. We identified and documented forecast evaluation methodologies with an open-source code base, WE-Validate, by setting up a benchmark of different forecast techniques. The significance of this benchmark exercise also lies in data sharing as well as knowledge sharing among collaborators. This exercise sheds light on how forecast evaluation can be transformed into lessons learned, thus leading to purposeful incentives for forecast improvements. Through this exercise and by demonstrating 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 put numerous data sets through the same data processing procedures and sets of evaluation metrics in WE-Validate and generate standardized numerical outputs. Thus, users can objectively compare forecasts of distinctive types, from relatively sleek ensemble means to highly fluctuating ramp forecasts. As a demonstration, this work synthesizes the results from processing multiple forecasts via WE-Validate. Additionally, users can use WE-Validate to select the ideal forecast providers [4,5]. 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 and transparent 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, when a minority of its members generate humongous forecast errors, an ensemble mean can display superb skills with RMSE while yielding a modest median absolute error. In that case, examining forecasting skills with only RMSE can lead to misinterpretation. Furthermore, analysts should spend time understanding the characteristics of the metrics they choose. For example, the PSS accounts for TN forecasts, but the SEDS does not. When forecasting rare events, we advise using SEDS for an impartial assessment. For a holistic forecast evaluation on wind ramp events, analysts should consider various ramp definitions of different magnitude and periods 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 relatively smoothed 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 lead to satisfactory results and do not yield the largest errors among all ensemble members at any time steps. The wisdom of the crowd prevails in time series forecasts because an ensemble mean moderates the differences and underscores the common features between ensemble members. Hence, when compared to observations at every time step, ensemble means can achieve favorable performance.

However, attributed to its smooth pattern, ensemble means tends not to perform well in ramp forecasts. In deterministic ramp forecast evaluations using the 2 × 2 contingency table, ensemble means become less skillful than many of their ensemble members, assuming the members have adequate ramp forecasting skills. As an alternative, a 50% voting system among individual ensemble members on indicating ramp forecasts often yield more correct ramp predictions than an ensemble mean, where the latter has a higher chance of missing forecast ramps. Given a voting scheme, a trade-off usually exists between more TPs with fewer FNs and larger FPs. A sensitive voting scheme can lead to a larger number of forecast ramps, which increases the chances of TPs and FPs. When missing extreme ramp events brings costly consequences, a sensitive voting scheme can be useful. In such circumstances, a forecast with a high POD and a high bias is favorable. With the aid of a performance diagram, we can visualize the performance of numerous forecasts as well as the trade-off between sensitive and insensitive voting schemes. Overall, in contrast to time series forecasting, the wisdom of the crowd carries a different meaning in ramp forecasting, in which members 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.

7. 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 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 ensemble means. We found that the multi-organization 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, unsurprisingly, the 2×2 contingency table reveals that the ensemble mean tends to miss predicting ramps because of its smooth pattern. To overcome this hurdle and take advantage of the information provided in an ensemble, we developed an arrangement in which individual ensemble members vote to detect ramps at a given time step. Such a voting scheme preserves the ramp forecasting skills of each ensemble member and therefore performs well in ramp identification, especially for ramps of large magnitudes. However, the downside of a sensitive voting scheme is risking more false alarms, whereas its likely benefit is more hits and fewer misses.

Looking forward, we envision that the wind energy industry will acknowledge the advantages and benefits 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. For instance, implementing the framework discussed in Ref. [40] will further improve WE-Validate. We welcome community contribution to WE-Validate to refine the wind forecast validation process.

Data availability

WE-Validate and example data sets discussed in this manuscript are available at https://github.com/a2edap/WE-Validate.

CRediT authorship contribution statement

Joseph C.Y. Lee : Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Caroline Draxl :** Conceptualization, Methodology, Data curation, Writing – review & editing, Supervision. **Larry K. Berg :** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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 Uwe Wagner 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 (WFIP2) case study, and the data are publicly available at the Atmosphere to Electrons Data Archive and Portal (https://a2e.energy.gov/data). Atmosphere to Electrons (A2e), including the analysis conducted herein to support the WFIP2 wind speed and power forecasts evaluation, is supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy's Wind Energy Technologies Office. Data from A2e projects are stored and disseminated via the A2e Data Archive and Portal at https://a2e.energy.gov.

We are grateful to our collaborators who submitted data for the benchmark exercise. We thank the members of the International Energy Agency (IEA) Wind Task 36 for their valuable input during the conception phase of the benchmark exercise and WE-Validate. Special thanks go to Jesper Thiesen, Daniel Camacho, and their colleagues at ConWX, Helmut Frank and his colleagues at Deutscher Wetterdienst; John Zack and his colleagues at MESO, Inc.; Jana Fischereit, Gregor Giebel, Andrea Hahmann, Marc Imberger, and Xiaoli Guo Larsén at the Department of Wind Energy at the Technical University of Denmark; and Corinna Möhrlen and her colleagues at WEPROG. We also thank the communications team at the National Renewable Energy Laboratory for editing this paper, including Sheri Anstedt, Amy Brice, and Laura Carter. Finally, we thank Will Shaw for his invaluable inputs and his immense contribution to this work. Our best wishes on Will's retirement.

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.

References

[1] R.J. Bessa, C. Möhrlen, V. Fundel, M. Siefert, J. Browell, S. Haglund El Gaidi, B.M. Hodge, U. Cali, G. Kariniotakis, Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry, Energies 10 (2017) 1402 https://doi.org/10.3390/EN10091402, 10 (2017) 1402.

[2] J.C.Y. Lee, M.J. Fields, An overview of wind-energy-production prediction bias, losses, and uncertainties, Wind Energy Sci. 6 (2021) 311–365, https://doi.org/10.5194/WES-6-311-2021.

[3] A. Craig, M. Optis, M.J. Fields, P. Moriarty, Uncertainty quantification in the analyses of operational wind power plant performance, J. Phys. Conf. 1037 (2018) 052021, https://doi.org/10.1088/1742-6596/1037/5/052021.

[4] C. Möhrlen, J. Zack, IEA Wind Expert Group Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions - Part 1: Forecast Solution Selection Process, 2019. https://iea-wind.org/wp-content/uploads/2021/04/IEAWindTask36-RecommendedPractice_Part1.pdf.

[5] C. Möhrlen, J. Zack, J. Lerner, IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions - Part 2: Designing and Executing Forecasting Benchmarks and Trials, 2019. https://iea-wind.org/wp-content/uploads/2021/02/IEAWindTask36-RecommendedPractice_Part2.pdf.

[6] C. Draxl, L.K. Berg, L. Bianco, T.A. Bonin, A. Choukulkar, A. Clifton, J.W. Cline, I.v. Djalalova, V. Ghate, E.P. Grimit, K. Holub, J.S. Kenyon, K. Lantz, C. Long, J.K. Lundquist, J. McCaa, K. McCaffrey, J.F. Newman, J.B. Olson, Y. Pichugina, J. Sharp, W.J. Shaw, N.H. Smith, M.D. Toy, The Verification and Validation Strategy

within the Second Wind Forecast Improvement Project (WFIP 2), Golden, CO, 2019 https://doi.org/NREL/TP-5000-72553.
[7] J.M. Wilczak, M. Stoelinga, L.K. Berg, J. Sharp, C. Draxl, K. McCaffrey, R.M. Banta, L. Bianco, I. Djalalova, J.K. Lundquist, P. Muradyan, A. Choukulkar, L. Leo, T. Bonin, Y. Pichugina, R. Eckman, C.N. Long, K. Lantz, R.P. Worsnop, J. Bickford, N. Bodini, D. Chand, A. Clifton, J. Cline, D.R. Cook, H.J.S. Fernando, K. Friedrich, R. Krishnamurthy, M. Marquis, J. McCaa, J.B. Olson, S. Otarola-Bustos, G. Scott, W.J. Shaw, S. Wharton, A.B. White, The Second wind forecast improvement project (WFIP2): observational field campaign, Bull. Am. Meteorol. Soc. 100 (2019) 1701–1723, https://doi.org/10.1175/BAMS-D-18-0035.1.
[8] W.J. Shaw, L.K. Berg, J. Cline, C. Draxl, I. Djalalova, E.P. Grimit, J.K. Lundquist, M. Marquis, J. McCaa, J.B. Olson, C. Sivaraman, J. Sharp, J.M. Wilczak, The Second wind forecast improvement project (WFIP2): general overview, Bull. Am. Meteorol. Soc. 100 (2019) 1687–1699, https://doi.org/10.1175/BAMS-D-18-0036.1.
[9] J.B. Olson, J.S. Kenyon, I. Djalalova, L. Bianco, D.D. Turner, Y. Pichugina, A. Choukulkar, M.D. Toy, J.M. Brown, W.M. Angevine, E. Akish, J.W. Bao, P. Jimenez, B. Kosovic, K.A. Lundquist, C. Draxl, J.K. Lundquist, J. McCaa, K. McCaffrey, K. Lantz, C. Long, J. Wilczak, R. Banta, M. Marquis, S. Redfern, L.K. Berg, W. Shaw, J. Cline, Improving wind energy forecasting through numerical weather prediction model development, Bull. Am. Meteorol. Soc. 100 (2019) 2201–2220, https://doi.org/10.1175/BAMS-D-18-0040.1.
[10] G. Giebel, W. Shaw, H. Frank, C. Draxl, P. Pinson, G. Kariniotakis, C. Möhrlen, Task Proposal Extension of Task 36 Forecasting for Wind Energy, Phase II, 2018.
[11] J. Reback, W. McKinney, jbrockmendel, J. van den Bossche, T. Augspurger, P. Cloud, gfyoung, Sinhrks, A. Klein, M. Roeschke, S. Hawkins, J. Tratner, C. She, W. Ayd, T. Petersen, M. Garcia, J. Schendel, A. Hayden, MomIsBestFriend, V. Jancauskas, P. Battiston, S. Seabold, chris-b1, h-vetinari, S. Hoyer, O. Wouter, alimcmaster1, K. Dong, C. Whelan, M. Mehyar, pandas-dev/pandas: Pandas 1.0.3 (v1.0.3), (2020). https://doi.org/10.5281/zenodo.3715232.
[12] L. Bianco, I.v. Djalalova, J.M. Wilczak, J. Cline, S. Calvert, E. Konopleva-Akish, C. Finley, J. Freedman, A wind energy ramp tool and metric for measuring the skill of numerical weather prediction models, Weather Forecast. 31 (2016) 1137–1156, https://doi.org/10.1175/WAF-D-15-0144.1.
[13] C.W. Hansen, W.F. Holmgren, A. Tuohy, J. Sharp, A.T. Lorenzo, L.J. Boeman, A. Golnas, The solar forecast arbiter: an open source evaluation framework for solar forecasting, Conf. Rec. IEEE Photovolt. Spec. Conf. (2019) 2452–2457, https://doi.org/10.1109/PVSC40753.2019.8980713.
[14] G. McCabe, J. Prestopnik, J. Opatz, J. Halley Gotway, T. Jensen, J. Vigh, M. Row, C. Kalb, H. Fisher, L. Goodrich, D. Adriaansen, M. Win-Gildenmeister, J. Frimel, L. Blank, T. Arbetter, The METplus Version 4.1.2 User's Guide, 2022. https://github.com/dtcenter/METplus/releases.
[15] M.W. Liemohn, A.D. Shane, A.R. Azari, A.K. Petersen, B.M. Swiger, A. Mukhopadhyay, RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics, J. Atmos. Sol. Terr. Phys. 218 (2021) 105624, https://doi.org/10.1016/J.JASTP.2021.105624.
[16] J.W. Messner, P. Pinson, J. Browell, M.B. Bjerregård, I. Schicker, Evaluation of wind power forecasts—an up-to-date view, Wind Energy 23 (2020) 1461–1481, https://doi.org/10.1002/WE.2497.
[17] J. Zhang, A. Florita, B.M. Hodge, S. Lu, H.F. Hamann, V. Banunarayanan, A.M. Brockway, A suite of metrics for assessing the performance of solar power forecasting, Sol. Energy 111 (2015) 157–175, https://doi.org/10.1016/J.SOLENER.2014.10.016.
[18] D.S. Wilks, Statistical Methods in the Atmospheric Sciences, Academic Press, Amsterdam, Netherlands, 2011.
[19] D. Chakraborty, H. Elzarka, Performance testing of energy models: are we using the right statistical metrics? J. Build. Perform. Simulat. 11 (2018) 433–448, https://doi.org/10.1080/19401493.2017.1387607.
[20] D. Entekhabi, R.H. Reichle, R.D. Koster, W.T. Crow, Performance metrics for soil moisture retrievals and application requirements, J. Hydrometeorol. 11 (2010) 832–840, https://doi.org/10.1175/2010JHM1223.1.
[21] S. Kim, H. Kim, A new metric of absolute percentage error for intermittent demand forecasts, Int. J. Forecast. 32 (2016) 669–679, https://doi.org/10.1016/J.IJFORECAST.2015.12.003.
[22] S.K. Morley, T.v. Brito, D.T. Welling, Measures of model performance based on the log accuracy ratio, Space Weather 16 (2018) 69–88, https://doi.org/10.1002/2017SW001669.
[23] C. Gallego-Castillo, A. Cuerva-Tejero, O. Lopez-Garcia, A review on the recent history of wind power ramp forecasting, Renew. Sustain. Energy Rev. 52 (2015) 1148–1157, https://doi.org/10.1016/J.RSER.2015.07.154.
[24] M. Cui, J. Zhang, C. Feng, A.R. Florita, Y. Sun, B.M. Hodge, Characterizing and analyzing ramping events in wind power, solar power, load, and netload, Renew. Energy 111 (2017) 227–244, https://doi.org/10.1016/J.RENENE.2017.04.005.
[25] J. Zhang, M. Cui, B.M. Hodge, A. Florita, J. Freedman, Ramp forecasting performance from improved short-term wind power forecasting over multiple spatial and temporal scales, Energy 122 (2017) 528–541, https://doi.org/10.1016/J.ENERGY.2017.01.104.
[26] Á. Hannesdóttir, M. Kelly, Detection and characterization of extreme wind speed ramps, Wind Energy Sci. 4 (2019) 385–396, https://doi.org/10.5194/WES-4-385-2019.
[27] B.R. Cheneka, S.J. Watson, S. Basu, A simple methodology to detect and quantify wind power ramps, Wind Energy Sci. 5 (2020) 1731–1741, https://doi.org/10.5194/WES-5-1731-2020.
[28] M. Dorado-Moreno, L. Cornejo-Bueno, P.A. Gutiérrez, L. Prieto, C. Hervás-Martínez, S. Salcedo-Sanz, Robust estimation of wind power ramp events with reservoir computing, Renew. Energy 111 (2017) 428–437, https://doi.org/10.1016/J.RENENE.2017.04.016.
[29] T. Ouyang, X. Zha, L. Qin, Y. Xiong, H. Huang, Model of selecting prediction window in ramps forecasting, Renew. Energy 108 (2017) 98–107, https://doi.org/10.1016/J.RENENE.2017.02.035.
[30] R. Sevlian, R. Rajagopal, Detection and statistics of wind power ramps, IEEE Trans. Power Syst. 28 (2013) 3610–3620, https://doi.org/10.1109/TPWRS.2013.2266378.
[31] C. Ferreira, J. Gama, L. Matias, A. Botterud, J. Wang, A Survey on Wind Power Ramp Forecasting, 2011 https://doi.org/10.2172/1008309, Argonne, IL (United States).
[32] P.J. Roebber, Visualizing multiple measures of forecast quality, Weather Forecast. 24 (2009) 601–608, https://doi.org/10.1175/2008WAF2222159.1.
[33] L.R. Barnes, D.M. Schultz, E.C. Gruntfest, M.H. Hayden, C.C. Benight, CORRIGENDUM: false alarm rate or false alarm ratio? Weather Forecast. 24 (2009) 1452–1454, https://doi.org/10.1175/2009WAF2222300.1.
[34] L.S. Gandin, A.H. Murphy, Equitable skill scores for categorical forecasts, Mon. Weather Rev. 120 (1992) 361–370 https://doi.org/10.1175/1520-0493, 1992) 120.
[35] R.J. Hogan, C.A.T. Ferro, I.T. Jolliffe, D.B. Stephenson, Equitability Revisited, Why the "equitable threat score" is not equitable, Weather Forecast. 25 (2010) 710–726, https://doi.org/10.1175/2009WAF2222350.1.
[36] R.J. Hogan, E.J. O'Connor, A.J. Illingworth, Verification of cloud-fraction forecasts, Q. J. R. Meteorol. Soc. 135 (2009) 1494–1511, https://doi.org/10.1002/QJ.481.
[37] C.A.T. Ferro, D.B. Stephenson, Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events, Weather Forecast. 26 (2011) 699–713, https://doi.org/10.1175/WAF-D-10-05030.1.
[38] J. Wilczak, C. Finley, J. Freedman, J. Cline, L. Bianco, J. Olson, I. Djalalova, L. Sheridan, M. Ahlstrom, J. Manobianco, J. Zack, J.R. Carley, S. Benjamin, R. Coulter, L.K. Berg, J. Mirocha, K. Clawson, E. Natenberg, M. Marquis, The wind forecast improvement project (WFIP): a public–private partnership addressing wind energy forecast needs, Bull. Am. Meteorol. Soc. 96 (2015) 1699–1718, https://doi.org/10.1175/BAMS-D-14-00107.1.
[39] C. Draxl, R.P. Worsnop, G. Xia, Y. Pichugina, D. Chand, J.K. Lundquist, J. Sharp, G. Wedam, J.M. Wilczak, L.K. Berg, Mountain waves can impact wind power generation, Wind Energy Sci. 6 (2021) 45–60, https://doi.org/10.5194/WES-6-45-2021.
[40] M.B. Bjerregård, J.K. Møller, H. Madsen, An introduction to multivariate probabilistic forecast evaluation, Energy and AI 4 (2021), https://doi.org/10.1016/J.EGYAI.2021.100058.