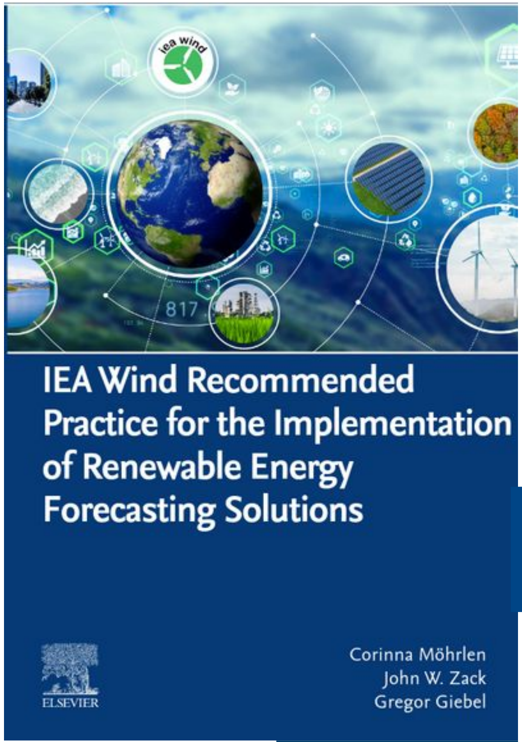
IEA Wind Task 51

IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions



Mathias Blicher Bjerregaard

Hands-on examples for the use of the guideline

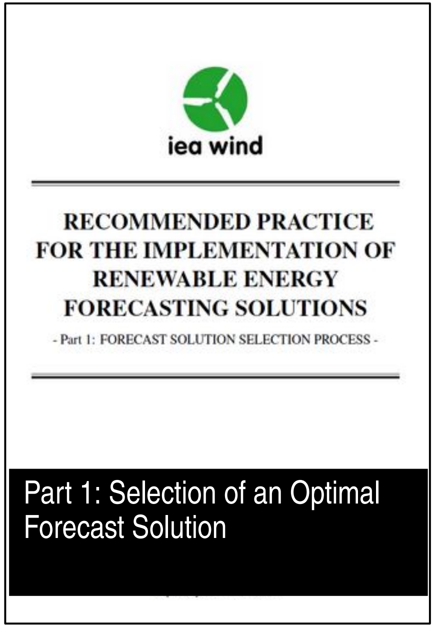
Windintegration Workshop
Kgs. Lyngby, Denmark, 25. Sept. *2023*

MESO ATMOSPHERIC RESEARCH FORECASTING · SIMULATIONS
WEPROG
DTU

Overview

- Background: IEA Wind Recommended Practice (RP) for the Implementation of Renewable Energy Forecasting Solutions
  - What it is
  - Where to get it
- Use Case Examples based on Recommendations
  - Wind speed evaluation at a Danish Coastal Site
  - Wind power evaluation at a substation in Ireland
  - Meteorological sensor performance assessment at a site in the German Bight (discussed in the paper – not included in this presentation)

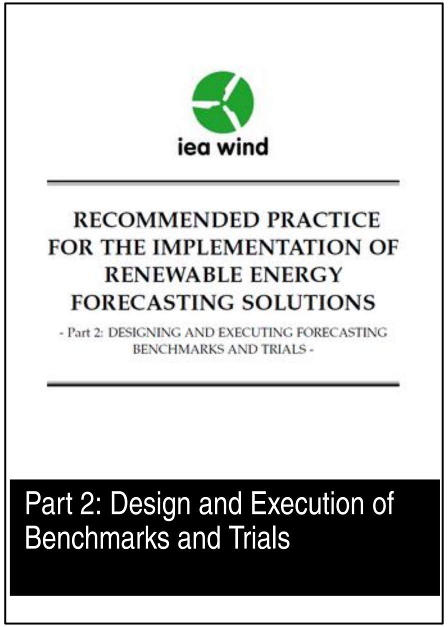
IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions: Set of 4 Parts


iea wind

RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

- Part 1: FORECAST SOLUTION SELECTION PROCESS -

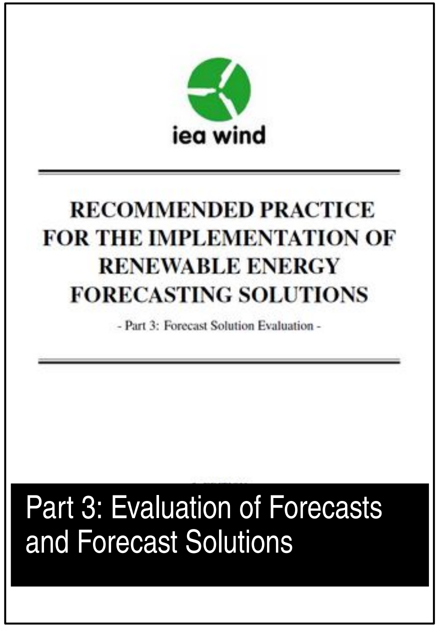
Part 1: Selection of an Optimal Forecast Solution


iea wind

RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

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

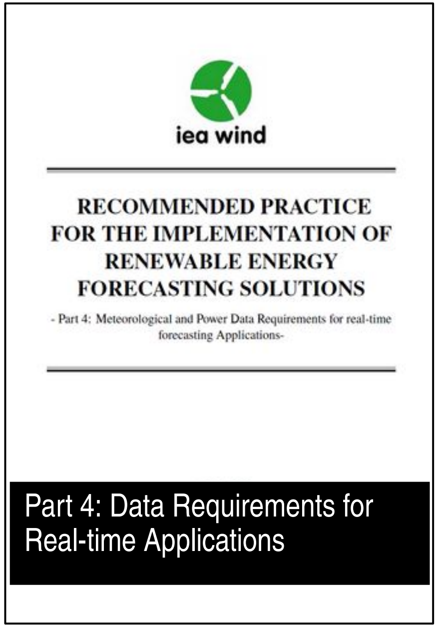
Part 2: Design and Execution of Benchmarks and Trials


iea wind

RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

- Part 3: Forecast Solution Evaluation -

Part 3: Evaluation of Forecasts and Forecast Solutions


iea wind

RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

- Part 4: Meteorological and Power Data Requirements for real-time forecasting Applications-

Part 4: Data Requirements for Real-time Applications

Video Introduction

Introduction: https://www.youtube.com/watch?v=XVO37hLE03M

IEA Wind Recommended Practice Book

Note

Elsevier Book

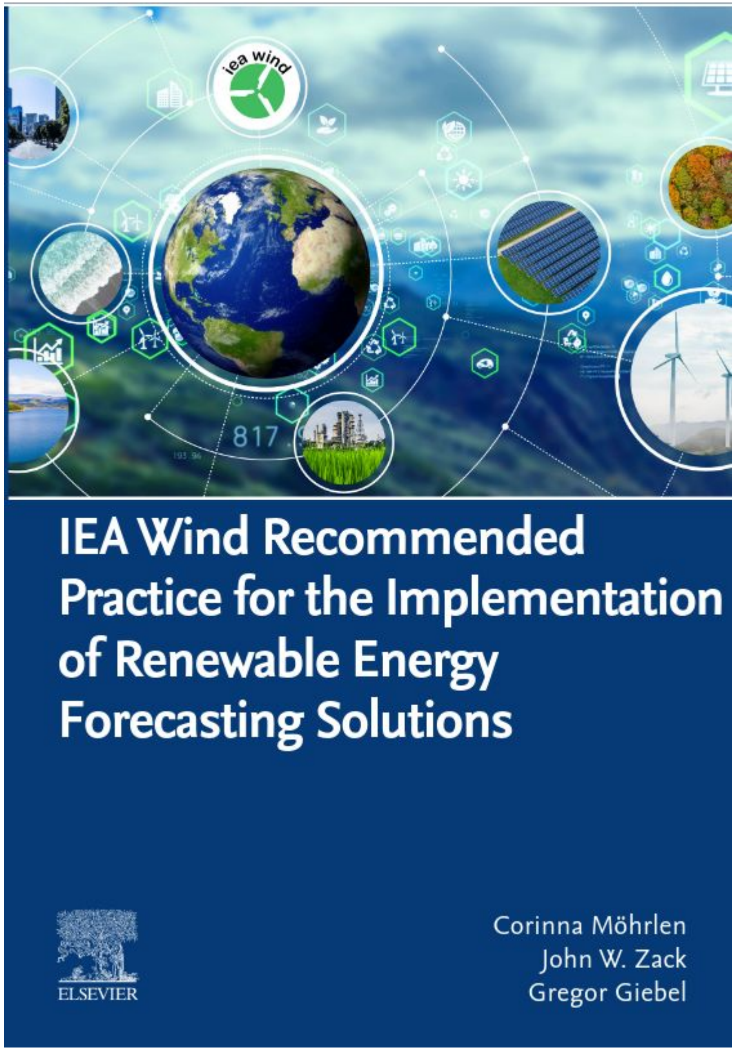
https://www.elsevier.com/books/iea-wind-recommended-practice-for-the-implementation-of-renewable-energy-forecasting-solutions/mohrlen/978-0-443-18681-3

Online OpenAccess:

https://www.sciencedirect.com/book/9780443186813/iea-wind-recommended-practice-for-the-implementation-of-renewable-energy-forecasting-solutions

IEA Wind Task 51 Information

iea-wind.org → Task 51 → Publications → Recommended Practice



IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecast Solutions

Application Areas for the Recommendations

1. System Operation, Balancing and Trading

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

2. Wind Turbine, Wind Farm and Solar Plant Operation and Monitoring

- Wind turbine and Power Plant Control
- Condition Monitoring

Companion Evaluation Software: “WE-validate-prob”

Assessment of forecasts with an R-package code

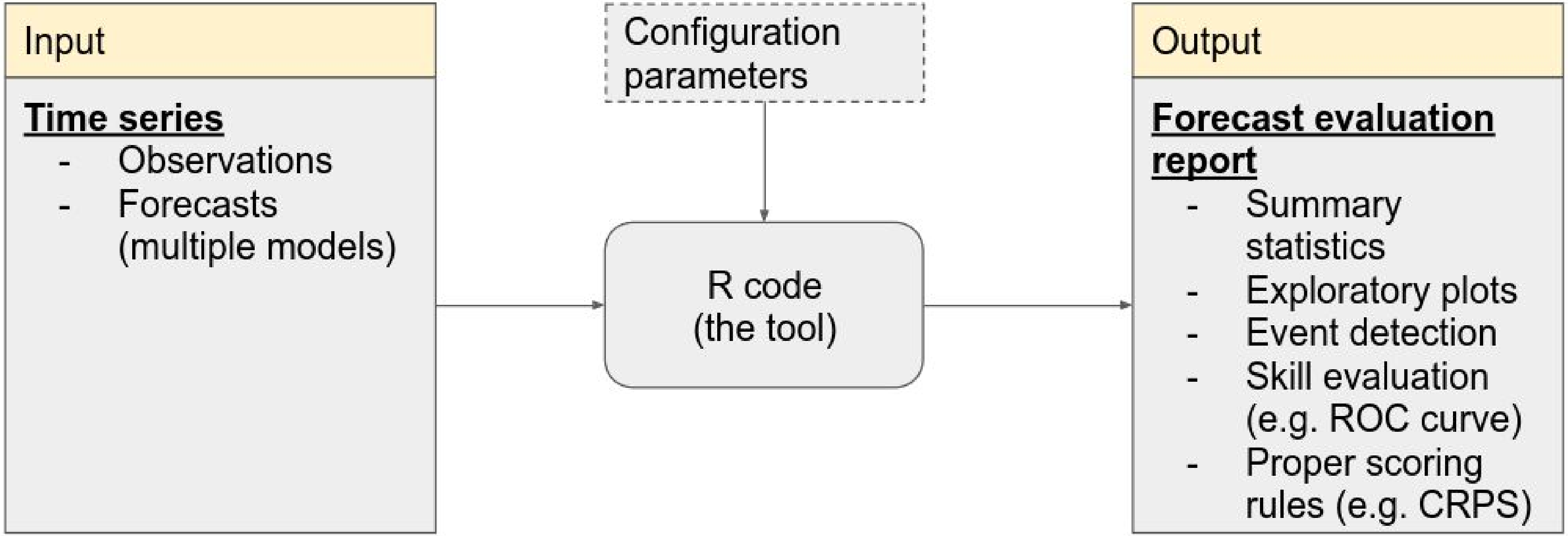
ACADEMIC PRESS

IEA Wind Recommended Practice for the Implementation of Renewable Energy Forecasting Solutions

Wind Energy Engineering

2023, Pages 321-322

Appendix G - Validation and verification code examples



Recommendation: Establish an Evaluation Framework

Key Components

Specify the forecast framework

(1) the forecast application
(2) the key forecast time frames
(3) a ranking of the importance of forecast performance attributes

Define the evaluation sample

(1) Choose a time period likely to produce a representative sample of relevant weather patterns
(2) Choose a sufficient and well-defined evaluation time frame (e.g. 3 months, 1 year, ...)

Quality control & delivery performance

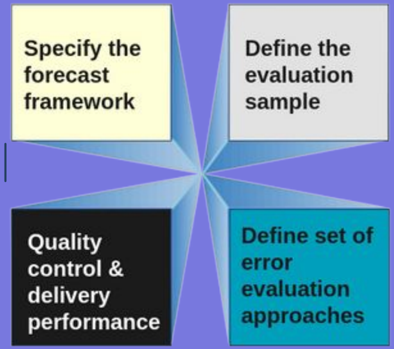
(1) Strategy to deal with missing or erroneous data & forecasts
(2) Specify evaluation criteria on delivery performance

Define set of error evaluation approaches

(1) visual inspection
(2) use of more specific metrics: SDE, SDBIAS, StDev, VAR, CORR
(3) use of histogram or box plot for evaluation of outliers
(4) use of contingency tables for specific event analysis
(5) use of improvement scores relative to a relevant reference forecast

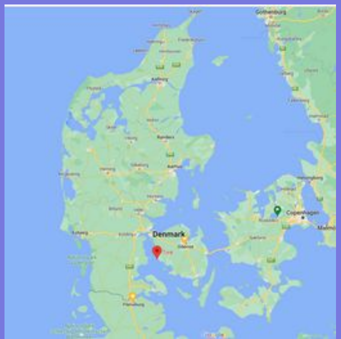
Example 1: Evaluation of Wind Speed at a Danish Coastal Site

Aim:
Verify the high resolution versus the low-resolution setup of an ensemble prediction system and evaluate improvement versus cost



Definition of the Sample:
Danish synoptic meteorological site:
South-west Funen “Assens”

- High-Resolution (HR):
  5km grid cells with 60 vertical levels
- Low resolution (LR):
  15km grid cells with 32 vertical levels

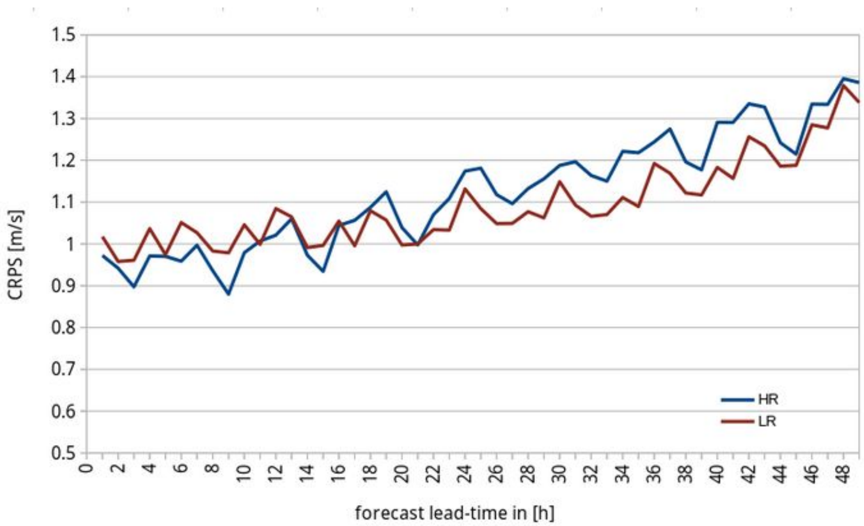


Evaluation Approach:
- CRPS
- CRPS lead-time dependency
- Reliability Diagram

Evaluation of Wind Speed at a Danish coastal site

Assessment of a high-resolution versus low resolution ensemble system

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



Result from Test 1:
High-resolution setup has only value in the first 12 hours

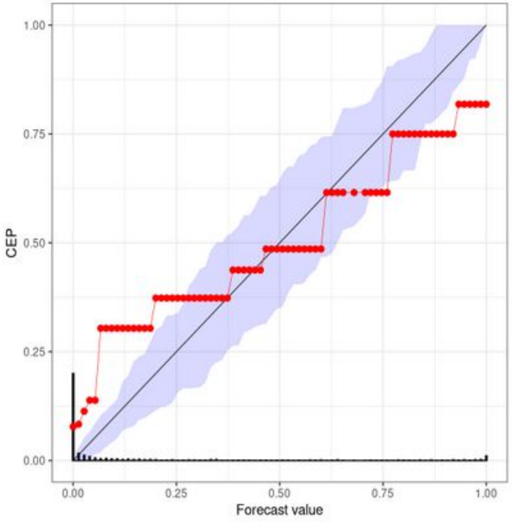
Conclusion:
High-resolution setup can be complementary in the intra-day…

Evaluation of Wind Speed at a Danish coastal site

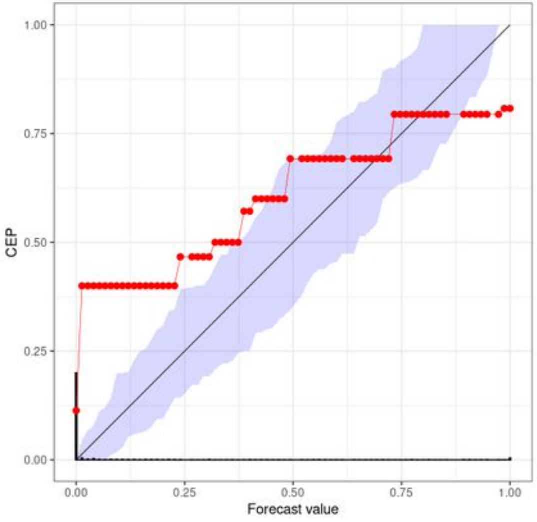
Assessment of a high-resolution versus low resolution ensemble system

Reliability is the degree to which the forecasted probabilities are in agreement with the outcome frequencies

HR-setup



LR-setup



Explanation of Score:
Reliability Diagram
X-axis: forecasted probabilities
Y-axis: conditional event probabilities (CEP) → the frequency of observed events given the specific forecast probability

Evaluation Criteria:

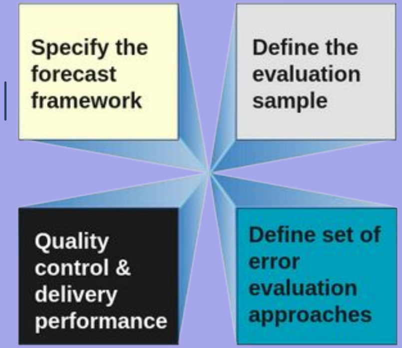
- Threshold: 5 (a minimum of 5 “positives” needed for an event)
- Forecast horizon: 6-11 hours
- Change in wind speed: 3m/s over a 3 hour window.

Result: tendency to lie on top of the diagonal for LR; Hr only in lower bins → indicates a negative BIAS and\or a slight mis-calibration

Conclusion: HR setup has a better balance between resolution and calibration, staying mostly within the blue 90% consistency band.

Example 2: Wind Power Evaluation at a Substation in the North-west of Ireland

Aim:
Verify forecasts from 2 different ensemble prediction systems and use a set of scores for decision-making regarding which of the forecasts provide better value



Definition of the Sample:
Sub station in North-west of Ireland:
A number of wind farms are feeding into the substation (wind farm cluster).
Forecast type:
Ramp forecasts
- High-Resolution (HR):
  5km grid cells with 60 vertical levels
- Low resolution (LR):
  15km grid cells with 32 vertical levels

Evaluation Approach:
- CRPS
- Brier Score
- ROC

Wind Power Evaluation at a Substation in the North-west of Ireland

Probabilistic Forecast Assessment of Ramping Events: **CRPS & Brier Scores**

CRPS score

overall performance of prob. forecast

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

No significance!

BRIER score

overall accuracy of a probabilistic *event* forecast

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

Large sensitivity to event choice!

Explanation of the score:

- CRPS is the probabilistic analogue to the Mean absolute error (MAE) for a deterministic forecast.
- Lower CRPS values indicate smaller error and therefore better performance.
- CRPS scores for each forecast over the 3-month test period

Explanation of the score:

- BS is the probabilistic analogue to mean squared error (MSE/RMSE) of deterministic forecast
- BS measures the mean squared difference (MSE/RMSE) between the forecasted probability (e.g., 0 to 1) and the actual outcome (e.g., 0 or 1).
- The BS values range between 0 and 1 with lower values indicating better performance.

Wind Power Evaluation at a Substation in the North-west of Ireland

Probabilistic Forecast Assessment of Ramping Events: **Brier Score Decomposition**

Decomposition of BRIER Scores

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

Explanation of the Scores:

- **Mean Score (MS)** measure the overall predictive event performance
- **Calibration/reliability (CAL)**: measures the agreement of forecasted probability with frequency of event occurrence given the forecasted probability (conditional event probability)
- **Discrimination/resolution (DSC/RES)**: measures the ability of forecasts to correctly distinguish differences in probabilities among the cases.
  → higher values contribute to lower BS, i.e. indicate better performance.
- **Uncertainty (UNC):** measures inherent uncertainty in the event and is related to the event frequency in the sample.
  → lower values contribute to lower BS, max. UNC for 50% events in sample

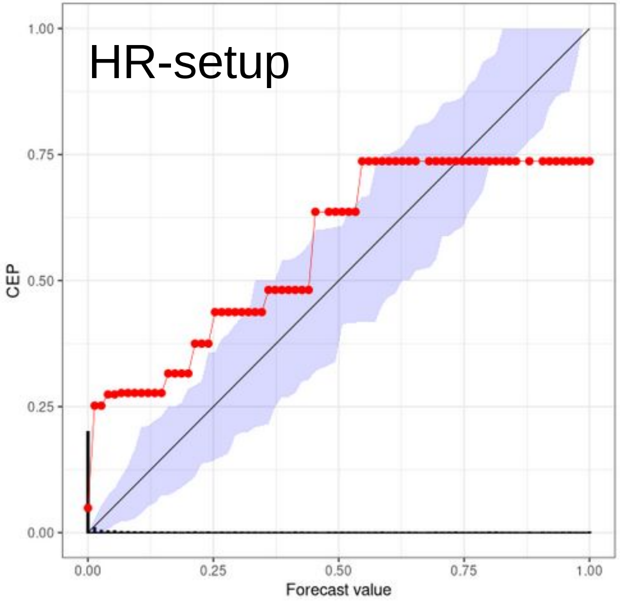
Result: The difference between HR and LR insignificant overall (MS), but quite significant for some components and sensitive to the thresholds and classifiers: the calibration (CAL) in the 40MW/3h class and the discrimination (DSC) in the 60MW/3h class is significantly better for the LR setup…

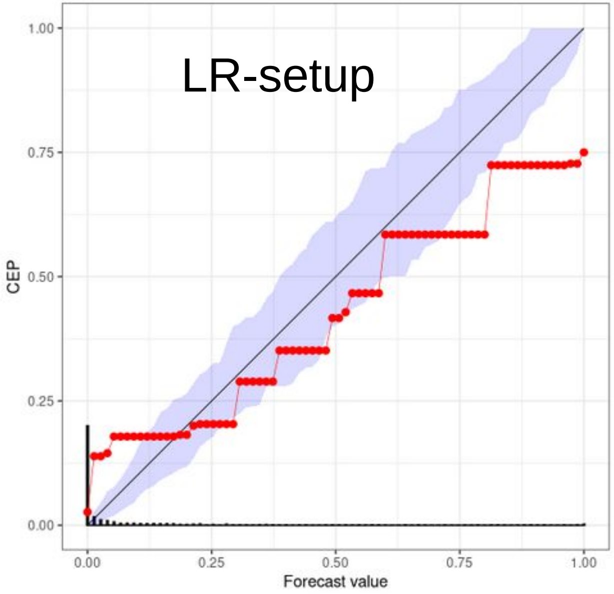
Conclusion: Decomposition of the Brier score is important, as it reveals differences in the forecast's skill related to distinguish events and to match occurrence with probabilities.

Wind Power Evaluation at a Substation in the North-west of Ireland

Probabilistic Forecast Assessment of Ramping Events: **Reliability Diagram**

Evaluation Criteria: Threshold: 5 - Forecast horizon: 6-11 hours - Change: 30MW over a 3 hour window.





Explanation of Plots:
X-axis: forecasted probabilities
Y-axis: conditional event probabilities (CEP) → frequency of observed events given the specific forecast probability
Band: 90% consistency band

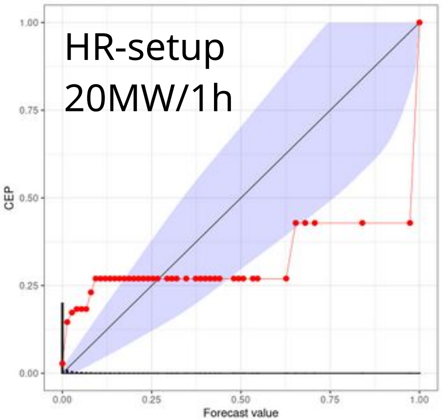
Result: tendency to lie on top the diagonal for HR; LR tendency to lie below diagonal
→ indicates a negative BIAS for LR and positive BIAS for HR … and/or a slight mis-calibration

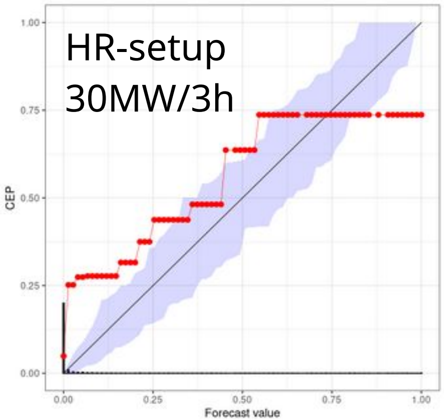
Conclusion: LR setup seems to be in better balance between resolution and calibration, staying mostly within the blue 90% consistency band – consistent with Brier score decomposition results….

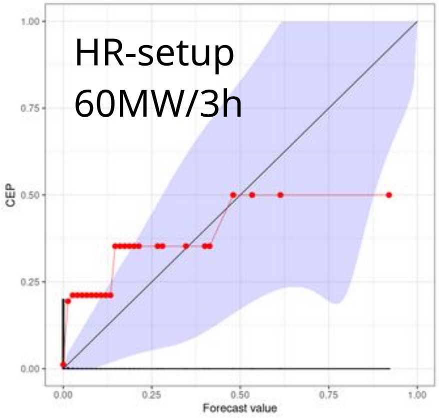
Wind Power Evaluation at a substation in the north-west of Ireland

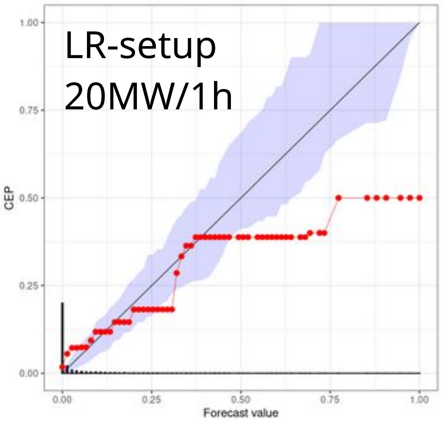
Probabilistic Forecast Assessment of forecasted Ramping Events: **Reliability Diagram**

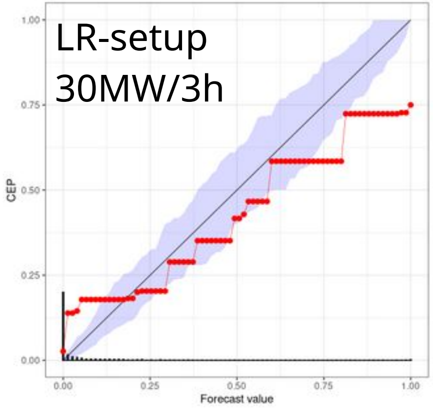
Demonstration of threshold selection sensitivity

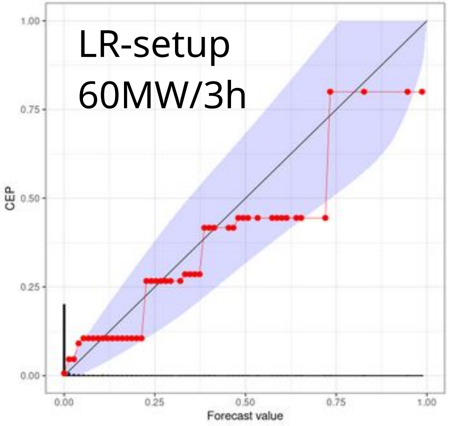












Evaluation of Wind Power at a substation in north-west of Ireland

Probabilistic Forecast Assessment of Ramping Events: **Continency table**

Contingency table +
HitRate (HR) and False Alarm rate (FAR)

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

Result:

LR forecasts have much higher number of “hits”

LR forecasts have much more “false alarms”

most extreme example of this pattern is for the 60MW/3hr threshold

→ requires to look into costs for misses versus false alarms…

Explanation of the Score:
The Contingency table lists: absolute number of "hits", "misses", "false alarms" and "correct negatives" in the forecast sample
lists the "hit rate" (HiR)
→ the hits per total number of forecasts
"false alarm rate" (FAR)
→ the false alarms per total number of forecasts.

Conclusion:

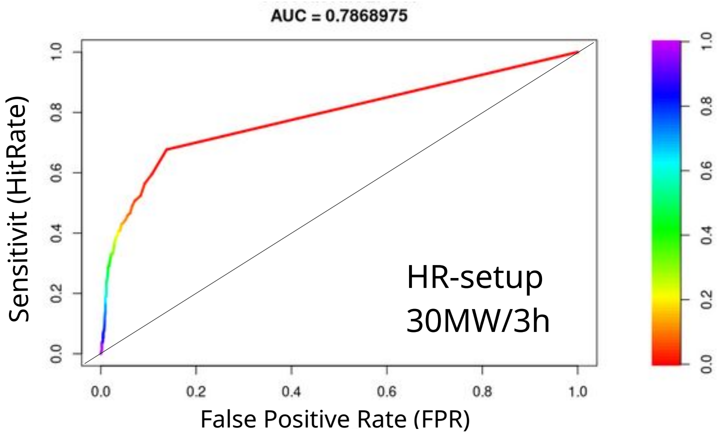
Beware of the threshold selection sensitivity in selection process and when analysing and evaluating the results

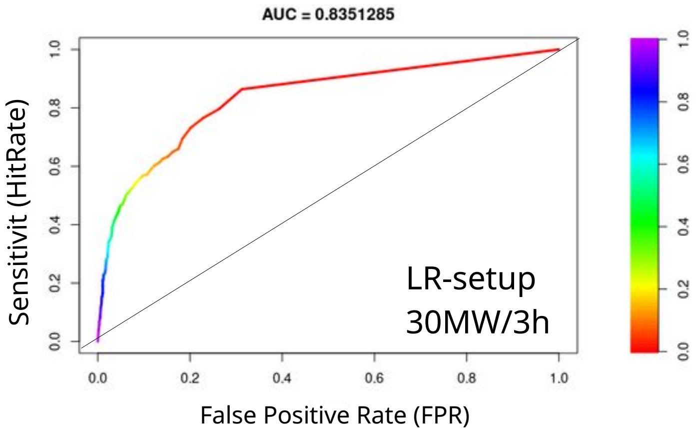
Fair evaluation comparison requires to provide the thresholds in advance

Wind Power Evaluation at a substation in the north-west of Ireland

Probabilistic Forecast Assessment of forecasted Ramping Events: **ROC Curve**

Receiver Operating Characteristics (ROC) curve measures the ability to discriminate between events and non-events and depicts the performance of forecasts at different probability thresholds





"Area Under the Curve" (AUC) for different ramping limits and time windows

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

Result:

Both forecast setups perform OK with a AUC > 0.7.

Slightly better, but little (insignificant) difference in the AUC scores for the LR forecasts

Explanation of the Score:

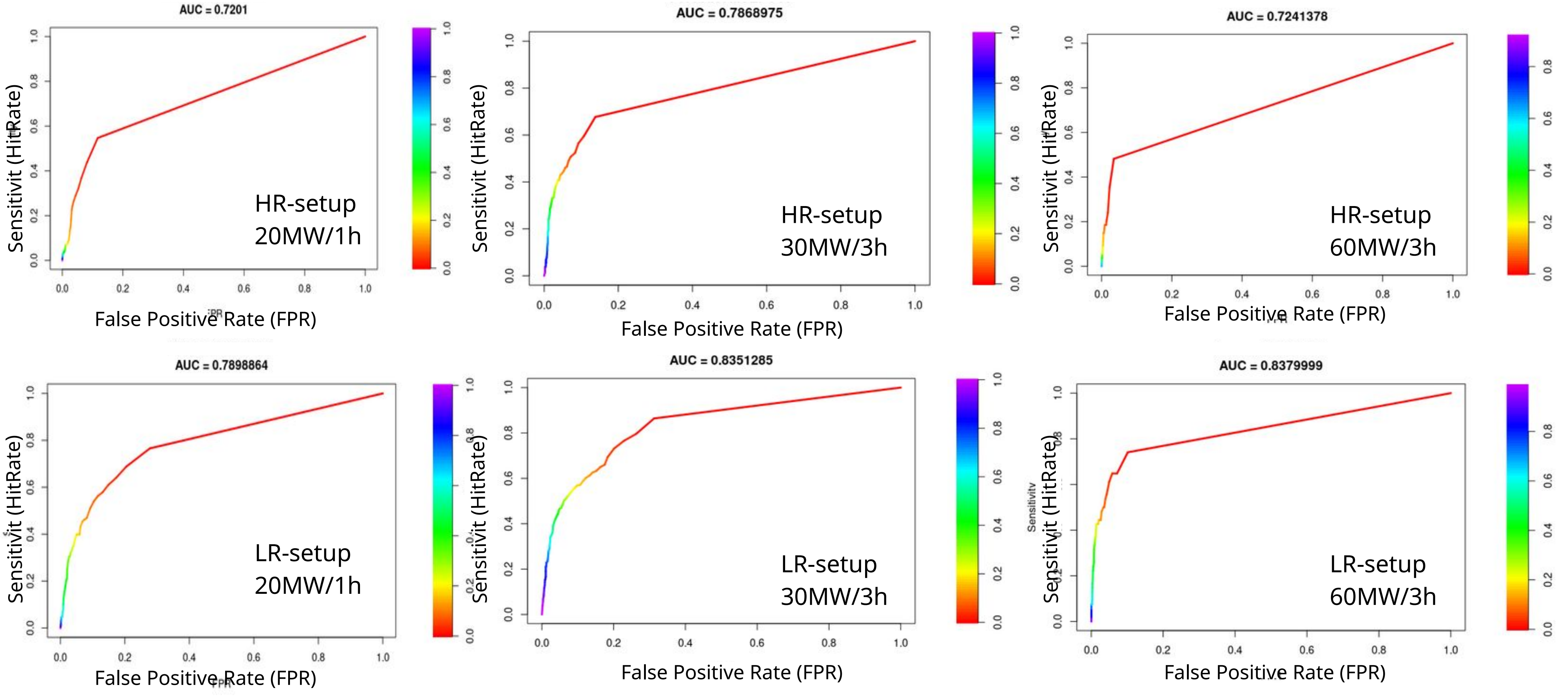
- The ROC curve ascends vertically at FAR=0.0 and horizontally at a sensitivity (hit rate) value of 1.0
- The color scale indicates classification thresholds yielding the points on the curve
- AUC= 1.0 for every forecast is a hit and no false alarms, 0.5 for random classifiers, i.e. forecasts with no skill (diagonal in graph)

Conclusion: the ROC curve confirms the results from the Brier Scores and indicates that the difference is not due to a mis-calibration.

Wind Power Evaluation at a substation in the north-west of Ireland

Probabilistic Forecast Assessment of forecasted Ramping Events: **ROC Curve**

Demonstration of threshold selection sensitivity



Evaluation of Wind Power at a Substation in North-west of Ireland

Probabilistic Forecast Assessment of Ramping Events: **Composite Performance Metric**

Assessment of the Forecast Error Scores:

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

- For the raw (unweighted) scoring, the high-resolution (HR) setup has a lower composite score (is "worse") than the low resolution (LR) setup
- If weights are applied according to the importance of each error metric for this application, the HR has a higher score indicating it may be a better choice for this application

See also recommendations in chapter 15 of IEA Wind Recommended Practice book

Lessons Learned and Take-away

Forecast Evaluation is subjective…
remember the 4 cornerstones for meaningful evaluation

- select carefully what you verify
- use more than one score/metric
- check consistency by applying various tests
- evaluate the significance of your results

THANK YOU FOR YOUR ATTENTION

Questions

Follow us:

Project webpage: http://iea-wind.org/task51

Publications: https://iea-wind.org/task51/task51-publications

RP-page: https://iea-wind.org/task51/task51-publications/task51-recommended-practices/

Contact us…

Presenter:

Dr. John W. Zack
WP2 Leader
MESO, Inc, USA
jzack@meso.com

Co-authors:

Dr. Corinna Möhrlen
WP3 Leader
WEPROG, DE & DK
com@weprog.com

Dr. Mathias Blicher B.
DTU Compute
Denmark

Dr. Gregor Giebel
Operating Agent
DTU Wind, Denmark
grgi@dtu.dk