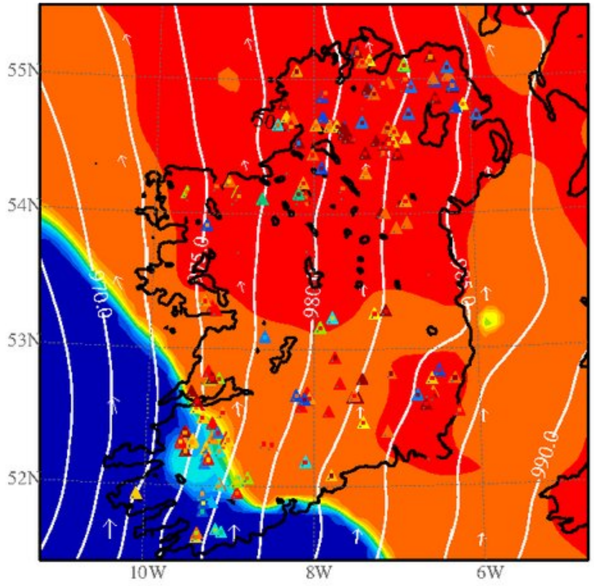
Minute-scale Forecasting Workshop

On the benefits and challenges of high-resolution ensemble forecasts for minute-scale ramping applications



Risø Campus, Denmark 10-11 April 2024

Dr. Corinna Möhrlen and Jess U. Jørgensen

About **WEPROG**

Check out our free weather WebAPP @ https://weather.weprog.com

First commercial established Windpower Forecast Vendor in Europe in 2003!

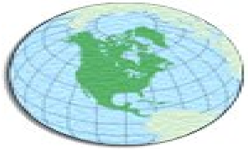
WEPROG's name is an abbreviation of **W**eather & **E**nergy **PROG**noses.

We provide **real-time ensemble forecasts** for a sustainable energy system

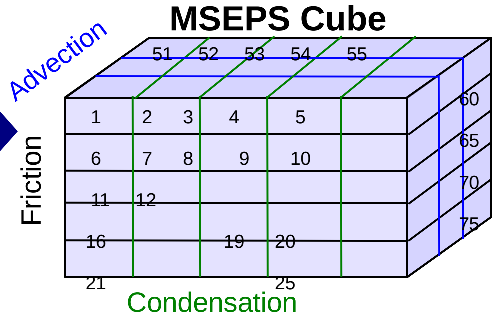
WEPROG operates a **continuous (24/7/365) forecast production.**

Real-time products are available in any region and all **continents**.

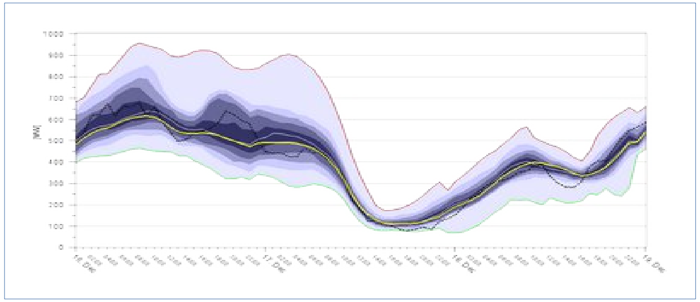
Global Analysis
(State Estimate)



75 independent weather forecasts

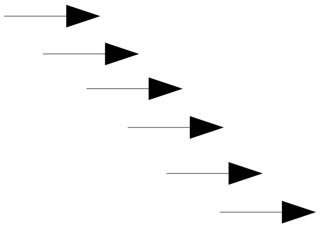


Probabilistic weather, wind & solar power forecasts

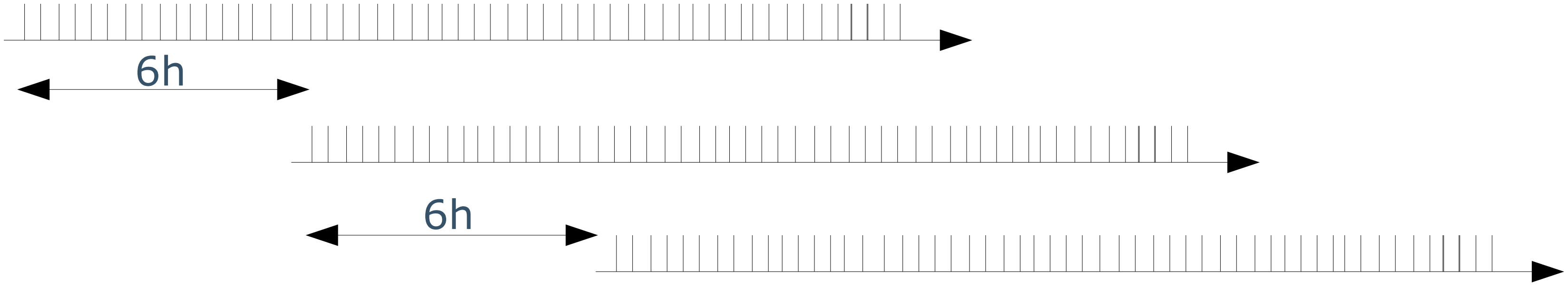


Minute-Scale Forecasting Applications

(1) Data-driven with high weight on persistence with a look ahead less than 2 hours with focus on precision frequently generated



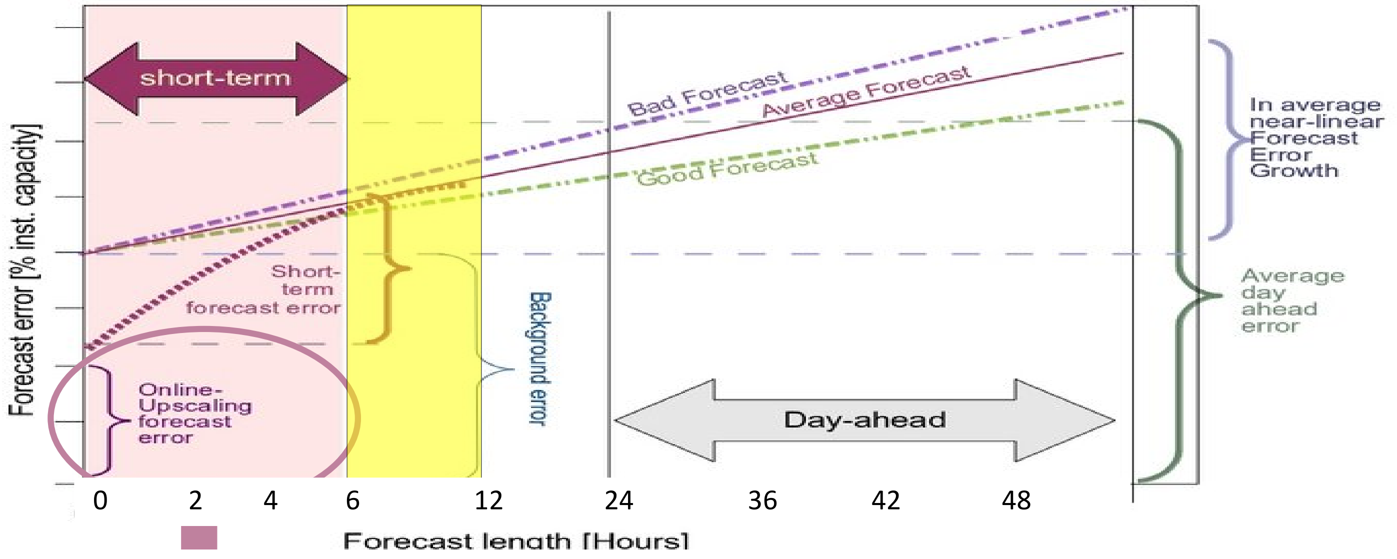
(2) Weather forecast driven risk based applications with a look-ahead of 2-48 hours with write out on minute-scale

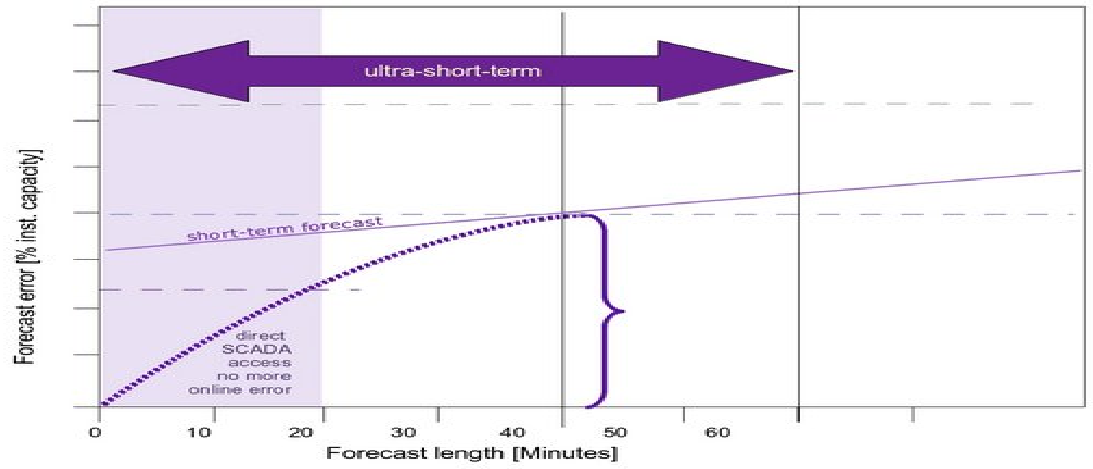


Move from Yesterday's time scales to Todays's requirements

TIME LENGTH of HOURS

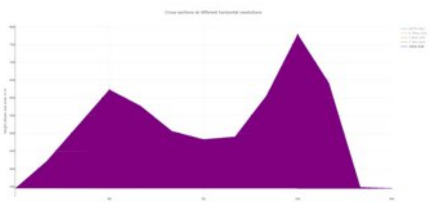
TIME LENGTH of MINUTES

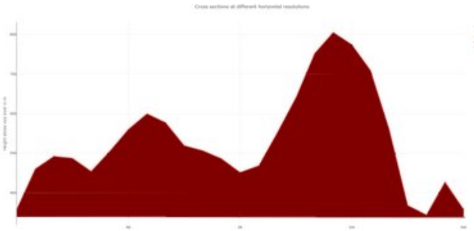




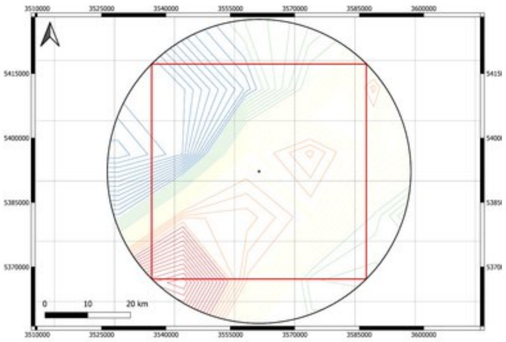
Visulalisations of different horizontal resolution in climate data for the numerical meso scale models

Profile of different model resolutions of the area around the WINSENT test field

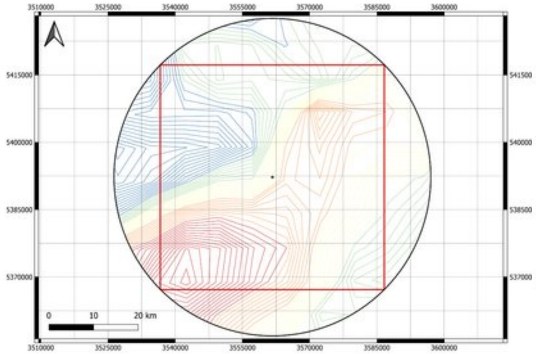




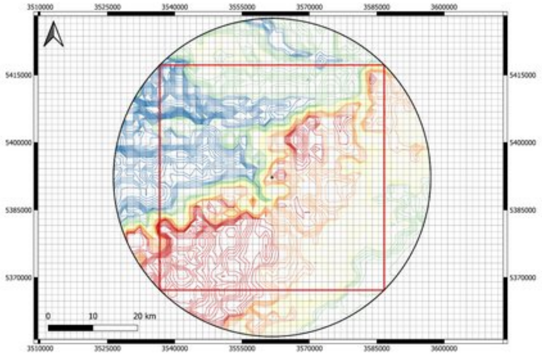
Horizontal profile in different resolutions of the area around the WINSENT test field



15km (operational)



7.5km (operational)



2.5km (research)

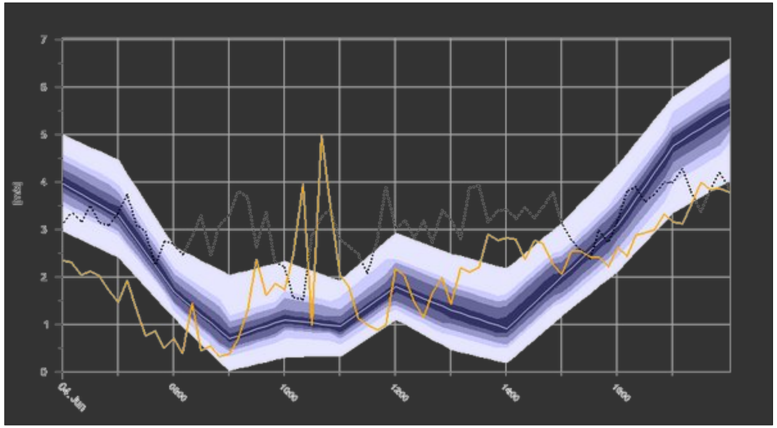
Minute Scale Forecasting with NWP Ensemble

CPU no problem any more

Post-processing no problem any more

Visually much closer to observations

Largest barrier in the past: Statistically high-resolution is often not better..



HR worse or NO significance!

Example: Wind Speed at 60m DTU Testsite Østerild			
Forecast	Bias	MAE	RMSE
LR best member	0.08	1.21	1.53
HR best member	0.11	1.19	1.52

Powerforecasts ca. 20 windfamrs in Ireland

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

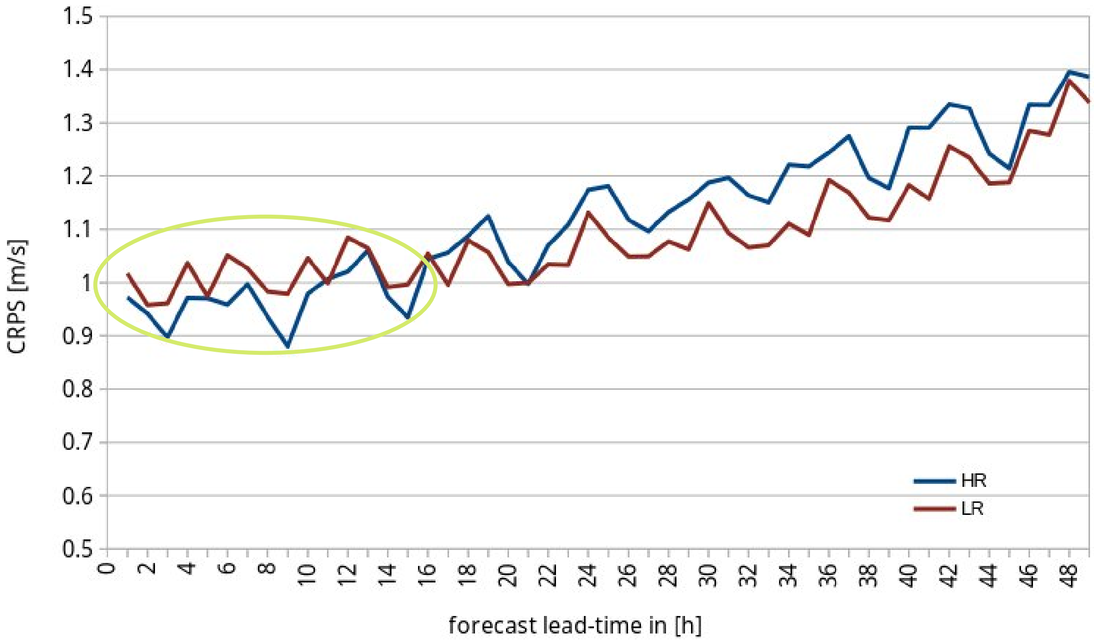
BRIER score example: accuracy of a probabilistic *event* forecast

Fore-cast	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

Example Evaluation of Wind Speed at a Danish coastal site

Assessment of a high-resolution versus low resolution ensemble system

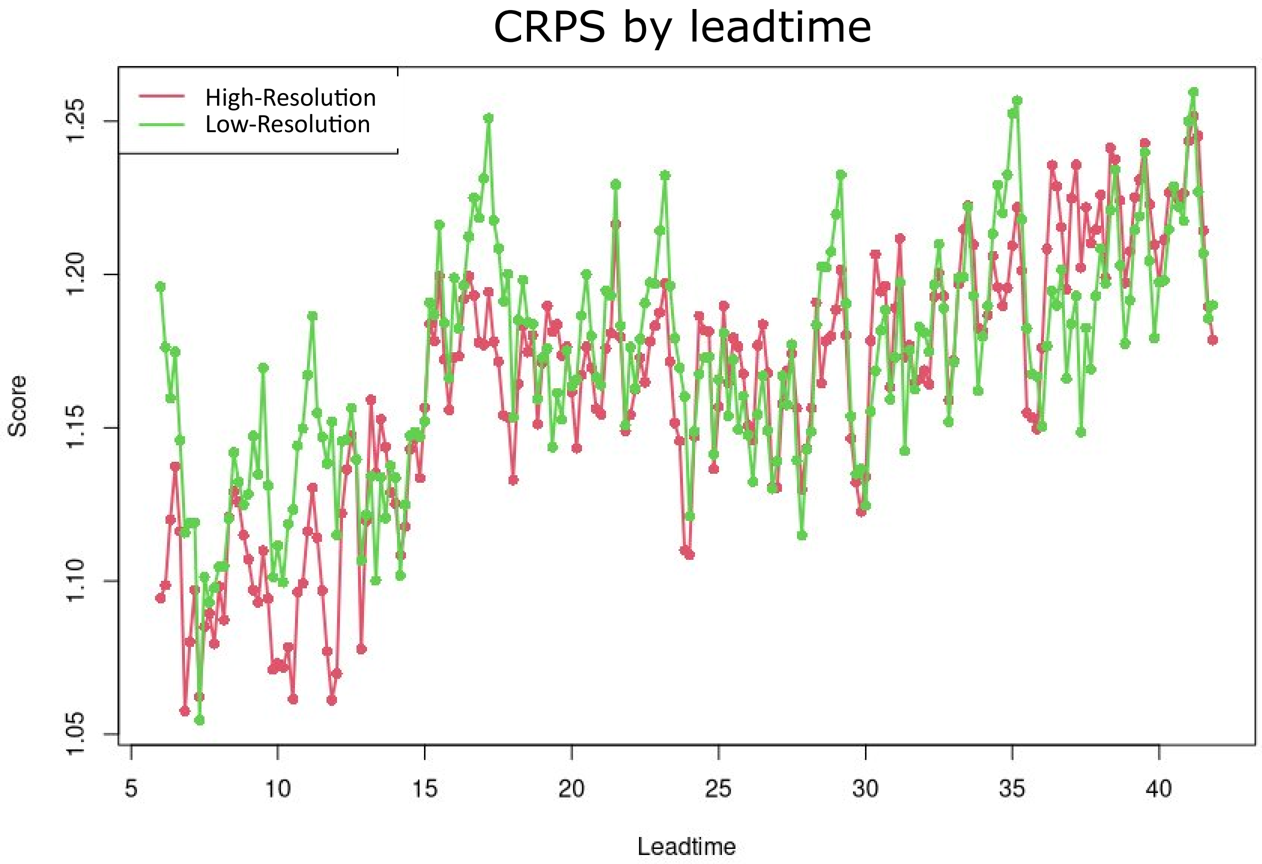
Forecast Type	CRPS	Improvement to Reference [%]
Reference	1.6635	
Lead-time	**6-11h**	
HR	1.140	-31.5
LR	1.159	-30.3
Lead-time	**0-48h**	
HR	1.1236	-32.5
LR	1.0925	-34.3



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...

Example Verification at WINSENT Testsite
CRPS: Wind Speed at 100m



High-Resolution worse or NO significance!

Forecast	CRPS
High-Resolution (f1)	1.163
Low-Resolution (f2)	1.171

Location: WINSENT Test site, Germany
Data period: Jan. - Nov. 2021
High-resolution ensemble: 5km
Low-resolution ensemble: 15km

Fundamental verification challenge

Back to the roots... NWP model development has for decades been based on

If you compare two model setups and they give equal objective verification scores, then take the one you believe is the better one, more future compatible, less costly and easier to maintain!!!

The dilemma mostly disfavours changes..

A new objective approach was needed that identifies detailed forecast skills for each model system of our target parameter.

We designed 24 tests with each 10 event-definitions to ensure that the answer "yes" implies an "event of interest" and which is difficult, but still realistic to forecast.

The correct answer is defined using the observation time series.

The forecast skill is quantified at different time-scale from the ratio of correct answers.

Observational Data for the Verification

Site: Stötten (WINSENT Testsite),
Southwest Germany

Period 1: Jan. - November 2021

1 Metmast with 4 measurement heights to 100m

Period 2: October 2023 – January 2024

Campaign in collaboration with ZSW and AQ Systems

1 AQ510 SODAR measureing up to 300m

1 Windscanner LIDAR measuring up to 1000m

2 Metmasts up to 100m

About the MSEPS Forecast data

Verification period: 145 intersecting days in 2021

Data Sampling and forecast horizon:
10-minute resolution from native model hour time resolution (<5min)
4 forecasts 48H per day 00,06,12,18 UTC
Only the 6-12 hours lead time is considered

LowResolution Setup – LR -
75 members and 32 levels – size 300 x 300

High-Resolution Setup – HR -
8 members and 60 levels - nested Domain in LR with 100 times more grid points
Acceleration at local scale can be ca. 200 times faster in HR than LR

Test methodologies to compare HR and LR

WEPROG

TEST 1: Skill measured using a deterministic answer per forecast

Symbol ∑ indicates a double sum $\sum_{time} \sum_{fc}$

a: Use all data

$$\mathbf{Skill}_{T1a} = \frac{\sum(True_{fc} True_{obs} + False_{fc} False_{obs})}{\sum(True_{obs} + False_{obs})}$$

b: Count only during events

$$\mathbf{Skill}_{T1b} = \frac{\sum True_{fc}}{\sum True_{obs}}$$

TEST 2: Skill measured using probability for each evaluation

Use only events

$$\mathbf{Skill}_{T2} = \frac{\sum_{time} \frac{100}{\#members} \sum_{fc} True_{fc} >= \mathbf{LTPCD}\ ?\ 1 : 0}{\sum_{time} True_{obs}}$$

Lower **T**hreshold **P**robability **C**riteria of **D**etection **(LTPCD)**

.

Test 2: Probabilistic answer using Ensembles

We determine **L**ower **T**hreshold **P**robability **C**riteria of **D**etection **(LTPCD)** for each model setup targeted the best detection.

The optimal **LTPCD** value is in the end a user choice determined by the cost ratio between success, miss and false alarms.

In this verification we select 3 different sample values to illustrate how the **LTPCD** value influence the result.

LTPCD for the HR setup

at least 1 member detects => 12.5% probability
at least 2 members detect => 25% probability
at least 4 members detect => 50% probability

LTPCD for LR is set lower to compensate for reduced model resolution

at least 3 members detect => 4.4% probability
at least 5 members detect => 6.6% probability
at least 10 members detect => 13.3% probability

Event-type definition at WINSENT Testsite for 100m wind speed

1. Does the wind speed **increase** at least 2.5m/s ?
2. Does the wind speed **increase** at least 2.0m/s ?
3. Does the wind speed **increase** at least 1.5m/s ?
4. Does the wind speed **increase** at least 1.0m/s ?
5. Does the wind speed **increase** at least 0.5m/s ?
6. Does the wind speed **drop** at least 0.5m/s ?
7. Does the wind speed **drop** at least 1.0m/s ?
8. Does the wind speed **drop** at least 1.5m/s ?
9. Does the wind speed **drop** at least 2.0m/s ?
10. Does the wind speed **drop** at least 2.5m/s ?

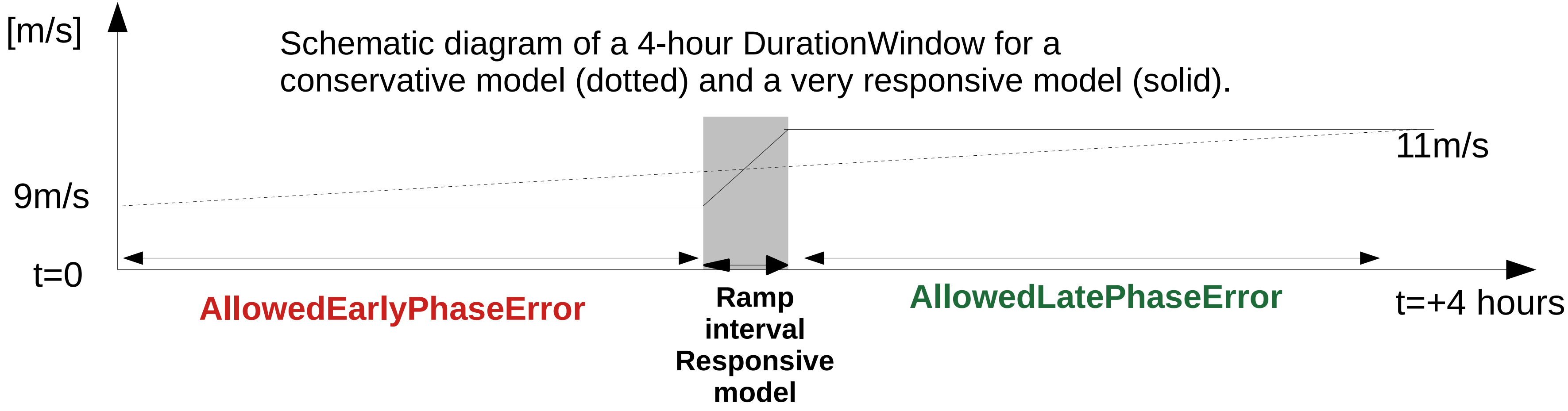
Event-detection was repeated every 10 min for 24 *DurationWindow* tests ranging from 10 min - 4h

→ YES answers from the observation data are considered events
→ NO answer is non-events.

25.030 tests got a YES, which is approximately 1 every 20min.

Duration Window and Ramp Interval concepts

We use 24 different DurationWindows ahead in time at every 10 minute interval. Within the DurationWindow, a change in wind speed can occur at a shorter interval (the RampInterval) together with an AllowedPhaseError.



The following relation applies

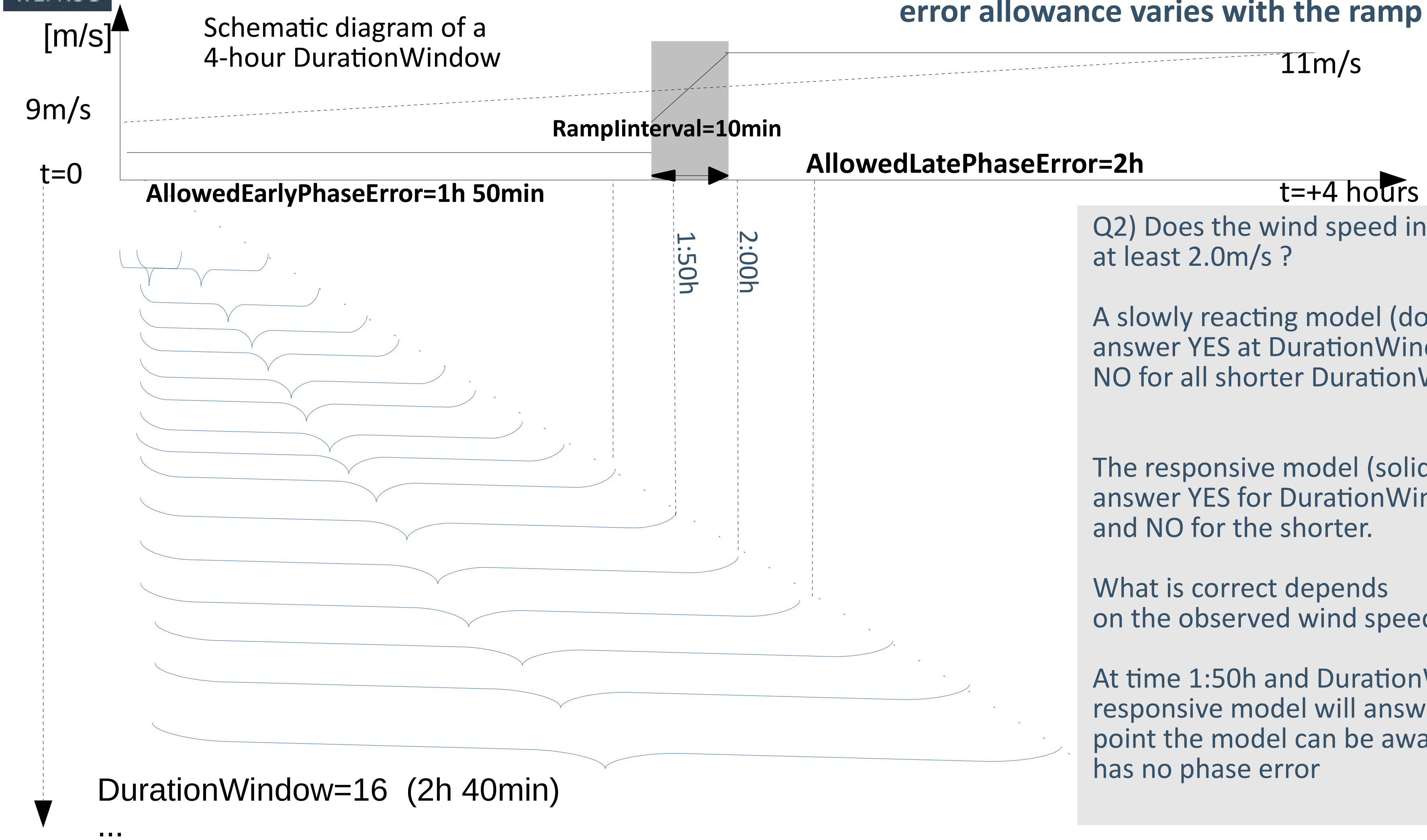
DurationWindow = RampInterval + AllowedEarlyPhaseError + AllowedLatePhaseError

The RampInterval cannot exceed the DurationWindow, but it is allowed to be shorter. If the RampInterval is short, then a correspondingly longer phase error tolerance is allowed.

The YES/NO answer is determined from difference between start and end point of the DurationWindow.

Ramp Interval definition

Increase of 2m/s + explanation of how the phase error allowance varies with the ramp window



Q2) Does the wind speed increase at least 2.0m/s ?

A slowly reacting model (dotted ramp) will answer YES at DurationWindow=4h, but NO for all shorter DurationWindow.

The responsive model (solid ramp) will answer YES for DurationWindow >= 2h and NO for the shorter.

What is correct depends on the observed wind speed change.

At time 1:50h and DurationWindow=1 the responsive model will answer YES. At that point the model can be awarded only if it has no phase error

Phase Error Allowance

The table below show hex numbers of how many 10min intervals a ramp may be shifted.
Rows represent the DurationWindow scale and columns represent the RampInterval.

RampInterval

AllowedPhaseError = (DurationWindow – RampInterval)/2

Time scale [10 min] / DurationWindow	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0																							
2	0	0																						
3	1	0	0																					
4	1	1	0	0																				
5	2	1	1	0	0																			
6	2	2	1	1	0	0																		
7	3	2	2	1	1	0	0																	
8	3	3	2	2	1	1	0	0																
9	4	3	3	2	2	1	1	0	0															
10	4	4	3	3	2	2	1	1	0	0														
11	5	4	4	3	3	2	2	1	1	0	0													
12	5	5	4	4	3	3	2	2	1	1	0	0												
13	6	5	5	4	4	3	3	2	2	1	1	0	0											
14	6	6	5	5	4	4	3	3	2	2	1	1	0	0										
15	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0									
16	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0								
17	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0							
18	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0						
19	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0					
20	9	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0				
21	a	9	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0			
22	a	a	9	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0		
23	b	a	a	9	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0	
24	b	b	a	a	9	9	8	8	7	7	6	6	5	5	4	4	3	3	2	2	1	1	0	0

Dark Green cells are the most difficult to forecast

Strongest Phase error allowance

All numbers represent number of 10 minute intervals

On robustness of the verification: number of Tests

Wind Speed Change	sign	Total Tests Count	Number of Events	Number of 10min scales
2.5	up	30829	763	1
2.0	up	30829	1163	4
1.5	up	30829	1803	14
1.0	up	30829	2953	56
0.5	up	30829	5077	262
0.5	down	30829	5788	262
1.0	down	30829	3409	68
1.5	down	30829	1984	19
2.0	down	30829	1258	8
2.5	down	30829	832	3
		308290	25030	697

Summary of the steps required to find skill differences

Preparation:

- Define 1-24 *Duration Windows* in 10-minute intervals (finest common granularity of obs/fc)
- Assemble 10-minute resolution of forecasts and observed values (hourly data is not enough!)
- Bias correct the forecasted wind speeds (same approach on each model system)
- Formulate ca. 10 questions on target variable changes from normal to extreme case
- Define **LTCPD threshold** for each ensemble member to be used in the probability calculation

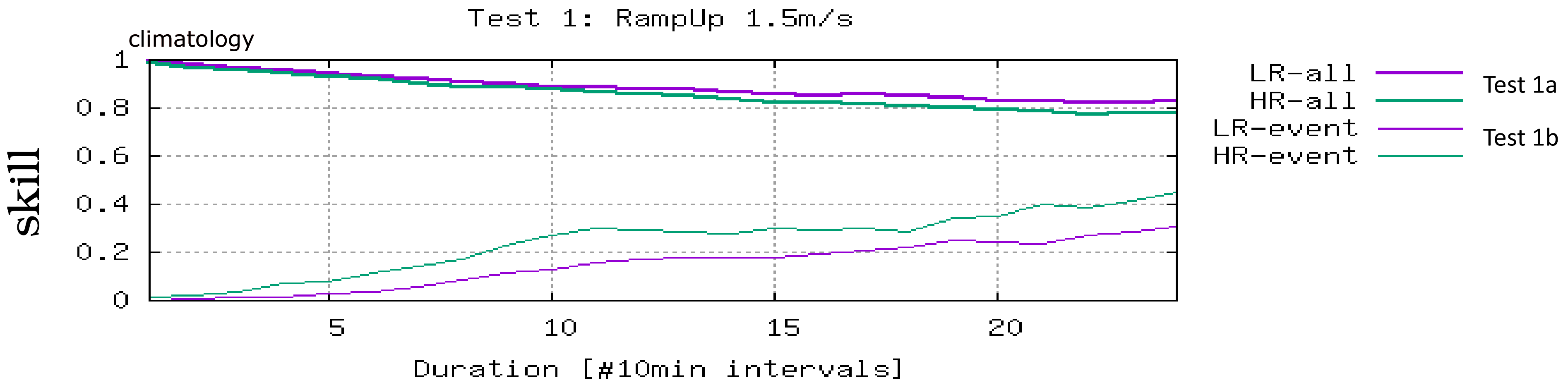
Processing:

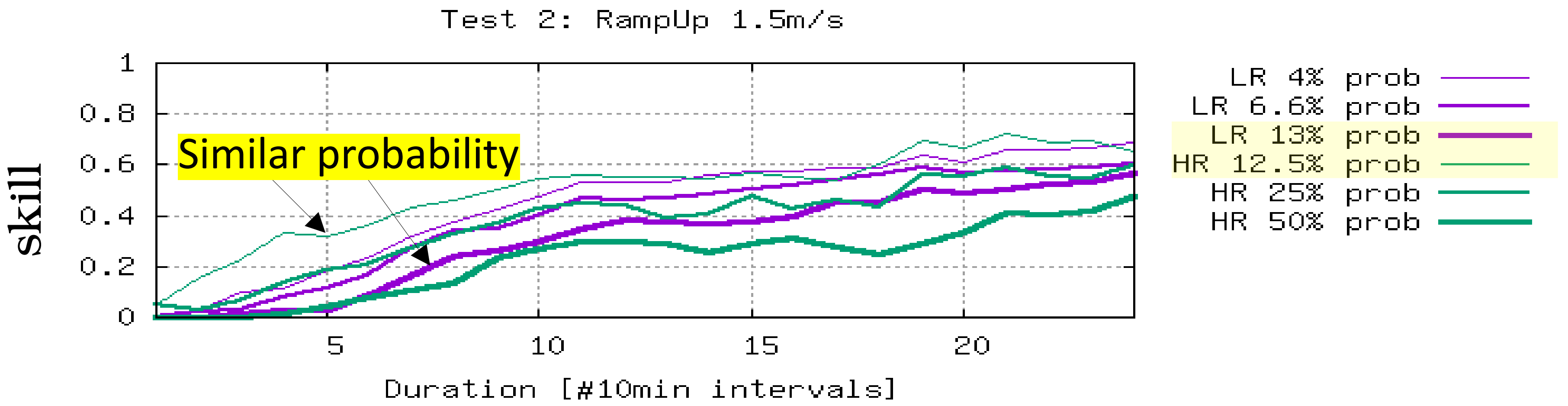
- Step in 10-minute resolution through the year and collect the 24 YES/NO values per model setup
- Calculate skill with 3 different formulas for each question
- Analyze results by question and assemble in table form

Final Evaluation:

- Perform winner count by row and column in the result table

Example Result plot for the 1.5m/s Ramp





Summary of Results from Graphs on Test 1a, Test 1b and Test 2

Wind Speed Change	Up or Down	Test 1a Winner of all Data	Test 1b Winner of all Events	Test 2 Winner per *Duration Window* interval: Short (1-6h)	Test 2 Winner per *Duration Window* interval: Long (18-24h)	
2.5	up	HR	HR	HR	HR	0 – 4
2.0	up	HR	HR	HR	HR	0 – 4
1.5	up	LR	HR	HR	EVEN	1 – 2
1.0	**up**	**LR**	**HR**	**HR**	**EVEN**	**1 – 2**
0.5	**up**	**LR**	**HR**	**HR**	**LR**	**2 – 2**
0.5	**down**	**HR**	**HR**	**HR**	**LR**	**1 – 3**
1.0	**down**	**HR**	**HR**	**HR**	**LR**	**1 – 3**
1.5	down	HR	HR	HR	EVEN	0 – 3
2.0	down	HR	HR	HR	HR	0 – 4
2.5	down	HR	HR	HR	HR	0 – 4
		3 / 7	0 / 10	0 / 10	3 / 4	6 – 31

… the “typical” barrier for change …

Verification Results from LR and HR comparison -- summary --

Note: The slope of the curves differ between Test 1a and Test 1b, because both model setups forecast non-events with high skill and events with lower skill.

Test 1a: **Good performance in at +/-0.5m/s at all time scales can be achieved with conservative model setup**.
→ High skill at short time scales occur because the amount of events is low compared to non events.
→ At longer time-scale skill is actually lower.

Test 1b: **low skill at short time scales and higher skill at longer time scales,**
because the longer time scales are equivalent larger spatial scales

Test 2: **probabilistic method provides useful and realistic skill comparison – even for a small ensemble !**
→ Moderate ramps have higher skill and HR and LR perform equal
→ The hit rate decreases for fast ramps, but is significantly higher for HR than LR
→ Fast ramps have some skill wtih HR, no skill with LR
→ The 12.5% **LTPCD** condition of HR is the most reliable of all 6 **LTPCD** conditions
→ LR detection is generally lower except for +/- 0.5m/s, although smaller **LTPCD** values are used

Result from the 4 comparison tests

HR setup has lower skill in modest ramp ups (0.5 and 1.0m/s) possibly due to false alarms. It is likely that HR generates meso-scale activity in this type of weather. If this type of activity does not happen, it could look like a weak point of HR.

LT shows higher skill for long time-scales and small ramps (0.5m/s) due to higher LTPCD value. If we compare **LTPCD**=12.5% for HR and **LTPCD**=13% for LR, then we find that HR and LR are even on the long time-scales.

The difficult part of Test 2 are the short time-scales without phase error allowance. Although HR is not great, it is outperforming LR on these time scales.

Forecast Skill

The verification methodology provides a sharp test to quantify forecast skill at the important risk related short time-scales.

Forecast Tuning

Ultimately this verification method can be used to further tune the forecasting toward predicting the need of the amount of primary reserve.

Lessons Learned and Take-aways

- event based probabilistic verification can provide a detailed time-scale dependent analysis of different model setups
- event based probabilistic approach is capable to compare fair performance & conditions of better or worse skills
- standard metrics are still relevant as independent tests and can reveal important information
- NWP models/Ensembles can provide skillful minute-scale forecasts beyond ST horizon and where there are no observations

Thank you for your attention !

Questions ...

Funded by:

WIKKI project funded by
Deutsche Bundesstiftung Umwelt (DBU) Germany
Project no. 37549/01

EARS4WindEnergy
Ensemble-based Approach utilizing a Refined SODAR for Wind Energy Applications

Innovation Fund Denmark

EUREKA Eurostar III Call 3 project
Project no. E2442 / Innovationsfonden 3109-00061B

Supported by:

WINSENT
Wind Science & Engineering Test Site in Complex Terrain

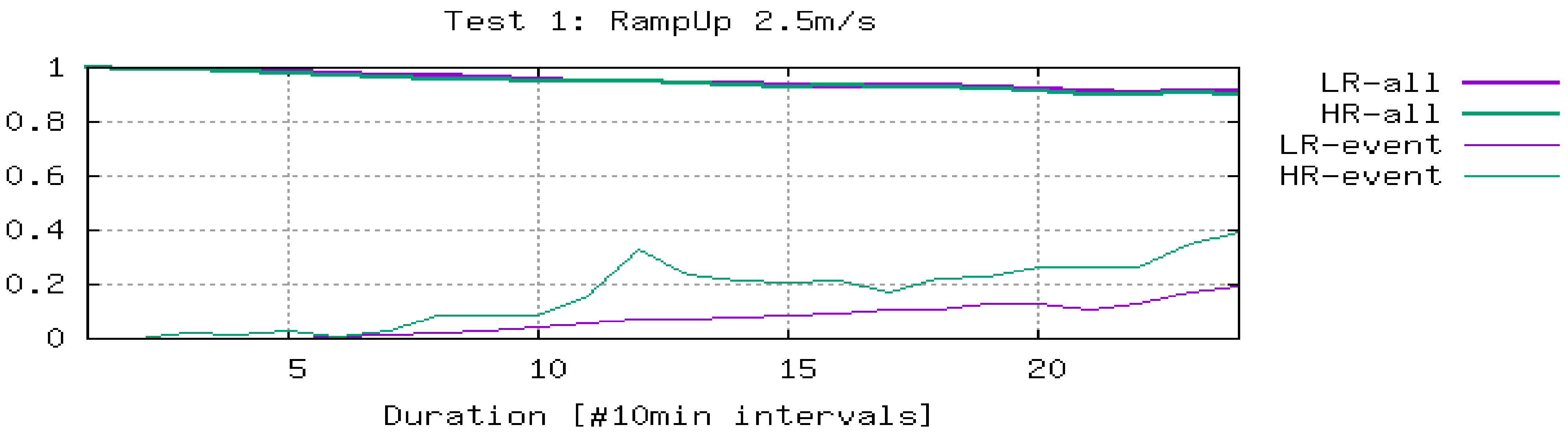
WinForS Consortium and WINSENT Testfield Coordinator:
ZSW Baden-Wurttemberg, Germany

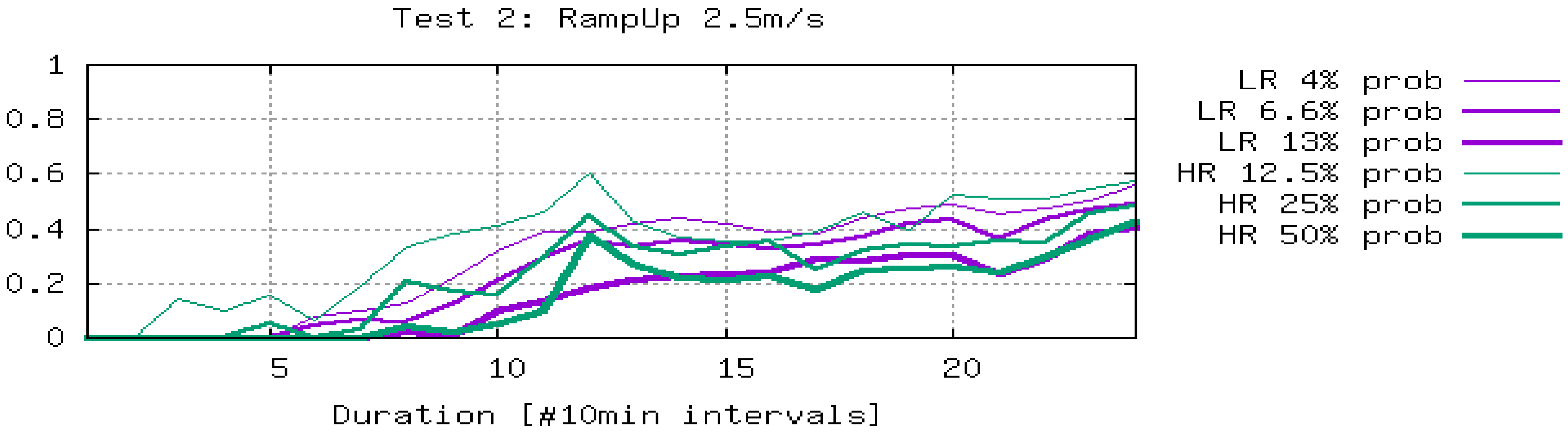
https://www.windfors.de/en/projects/test-site/winsent/

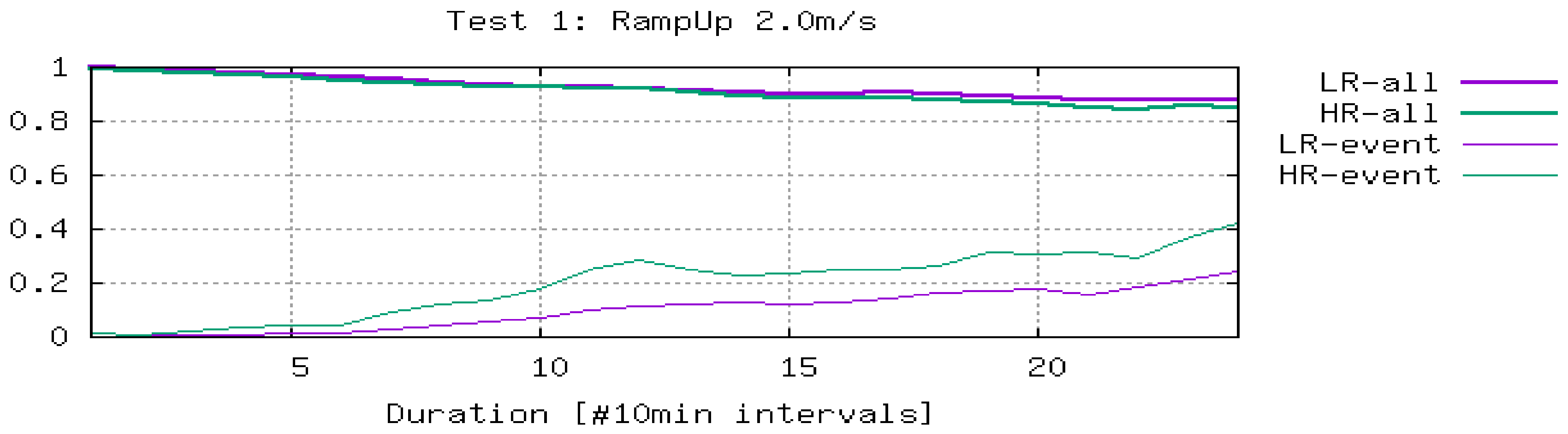
Contact:
Dr. Corinna Möhrlen
com@weprog.com

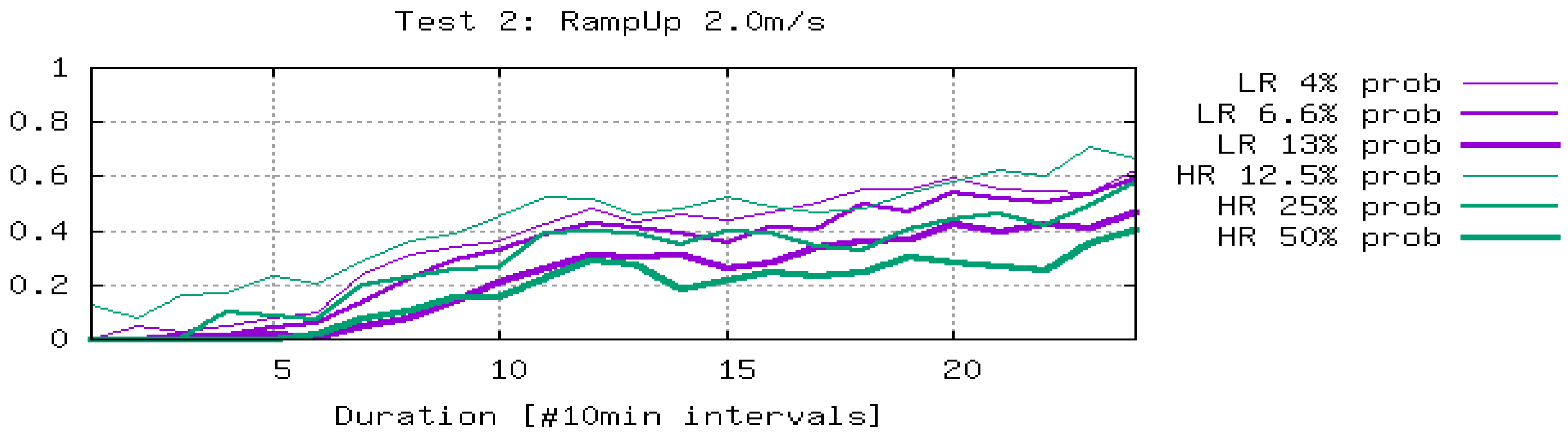
WEPROG
www.weprog.com
weather.wweprog.com

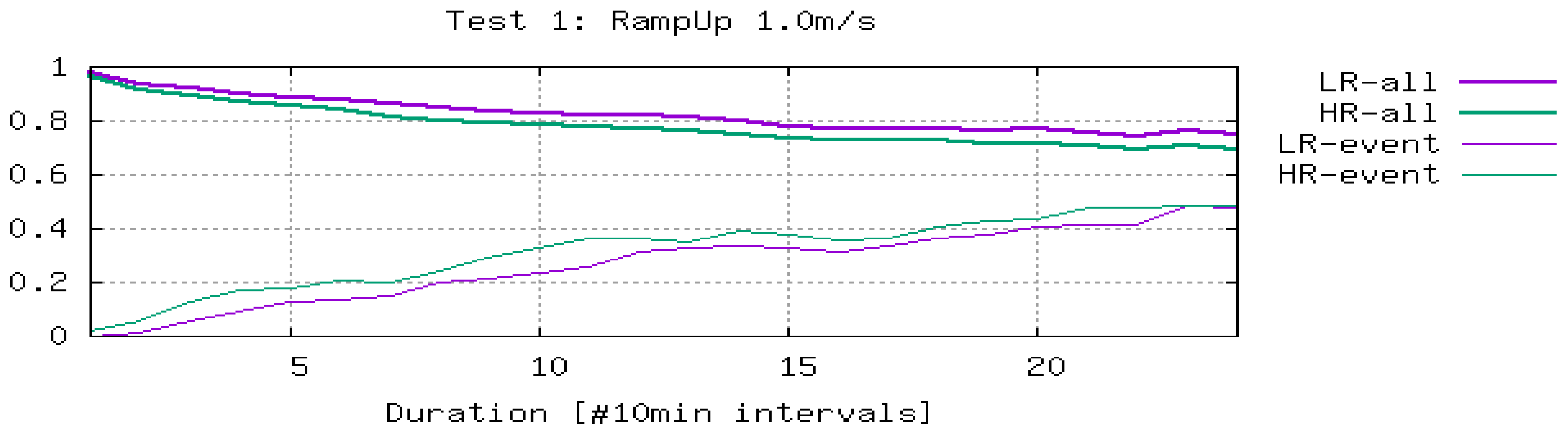
Result plots

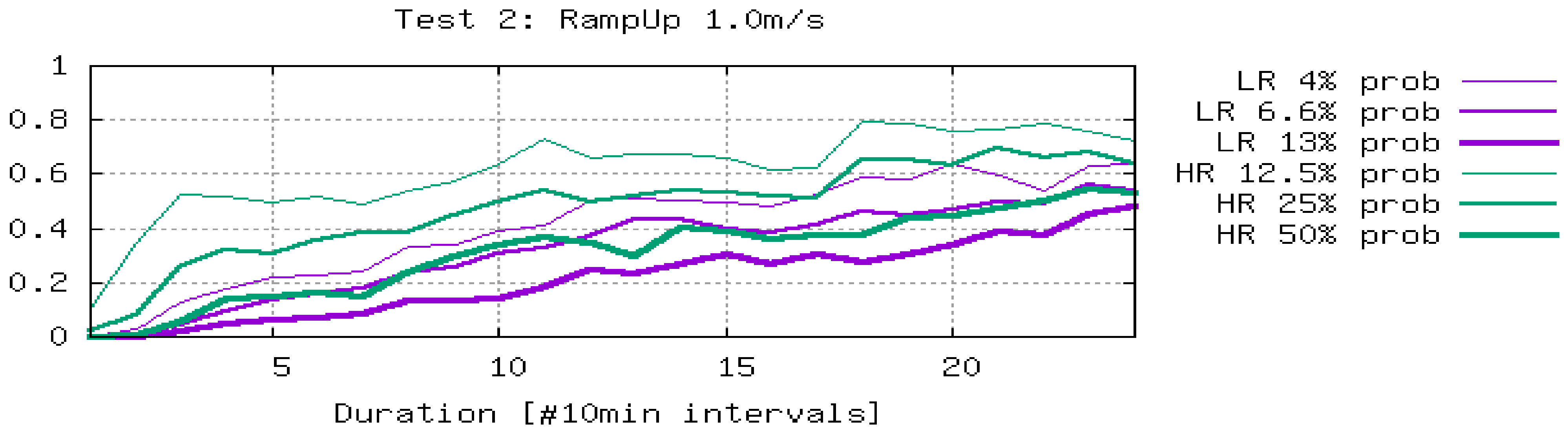

Test 1: RampUp 2.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event

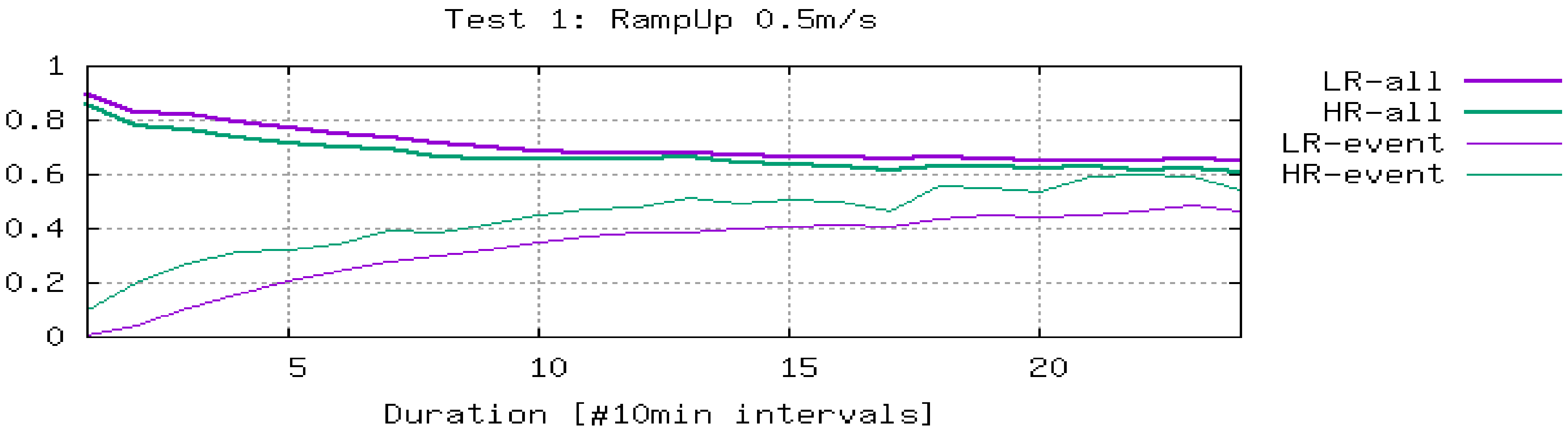

Test 2: RampUp 2.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob

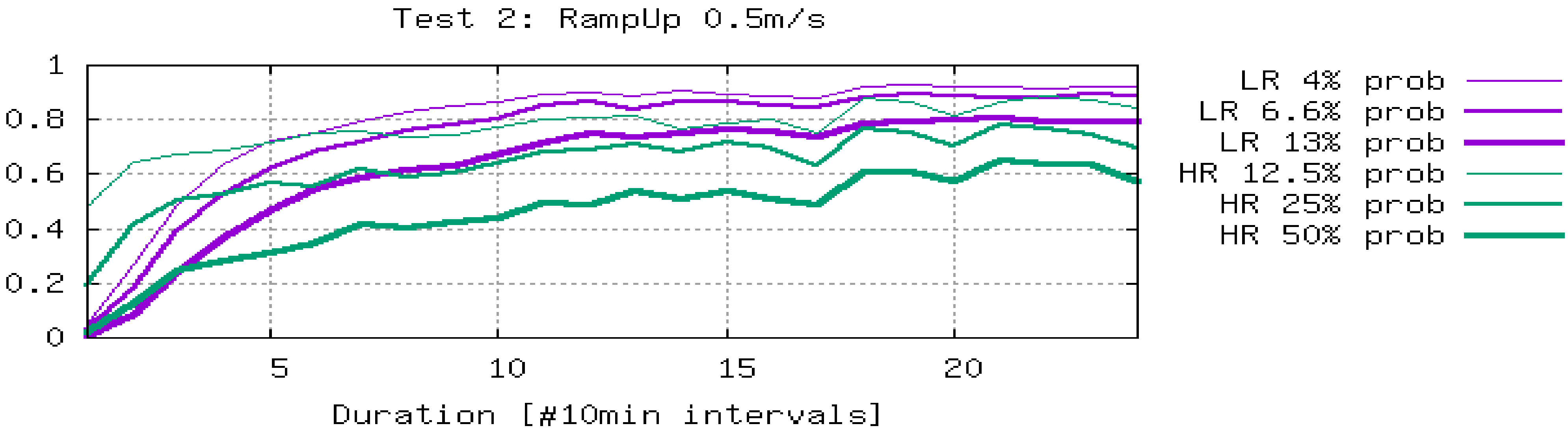

Test 1: RampUp 2.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event

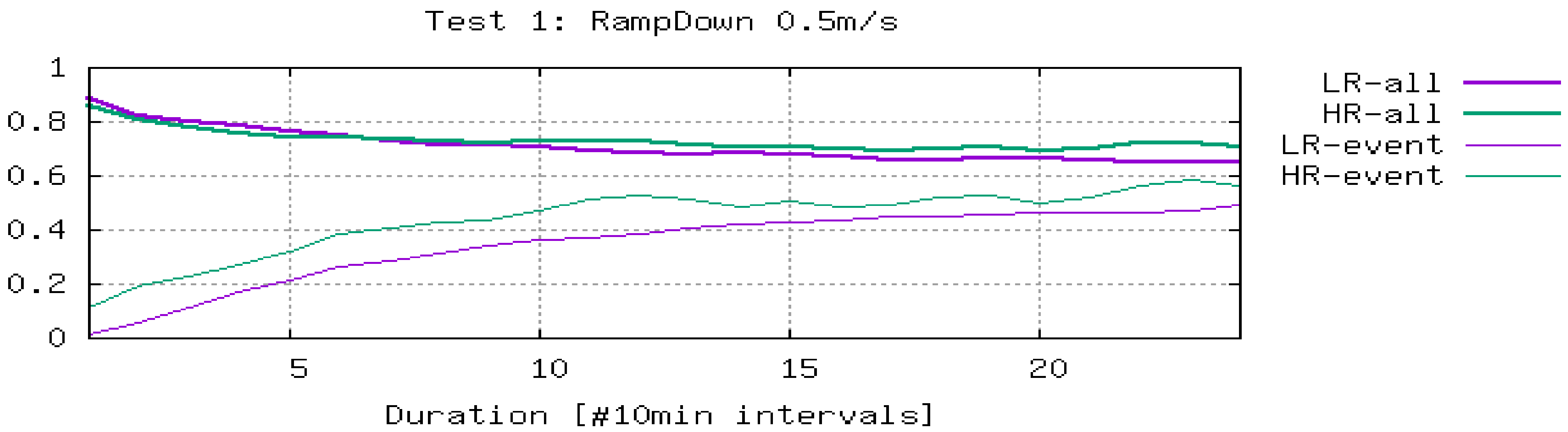

Test 2: RampUp 2.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob


Test 1: RampUp 1.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event

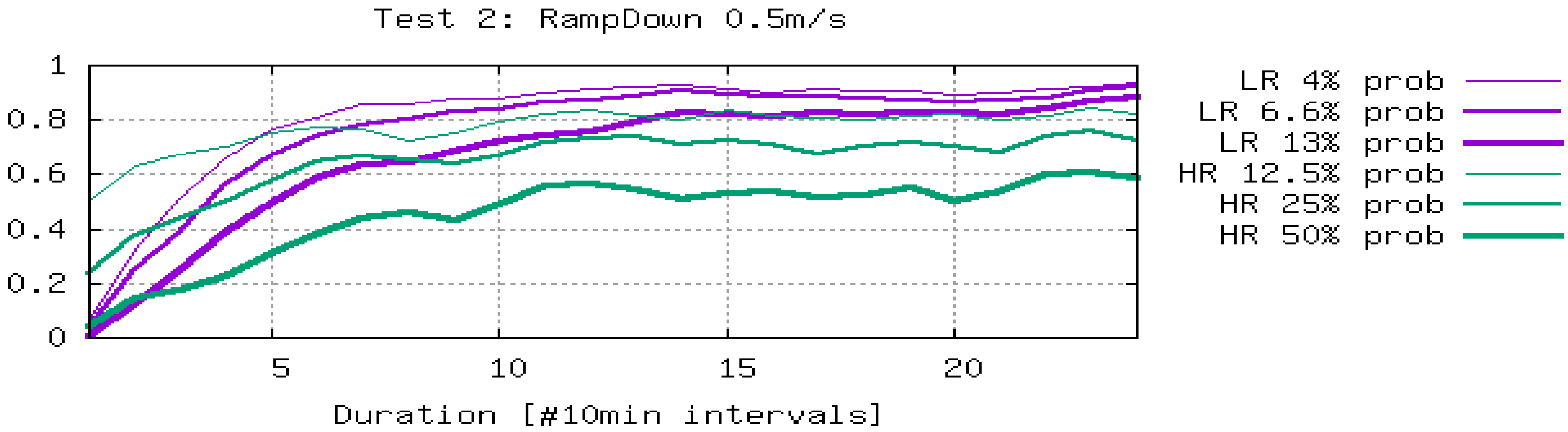

Test 2: RampUp 1.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob


Test 1: RampUp 0.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event

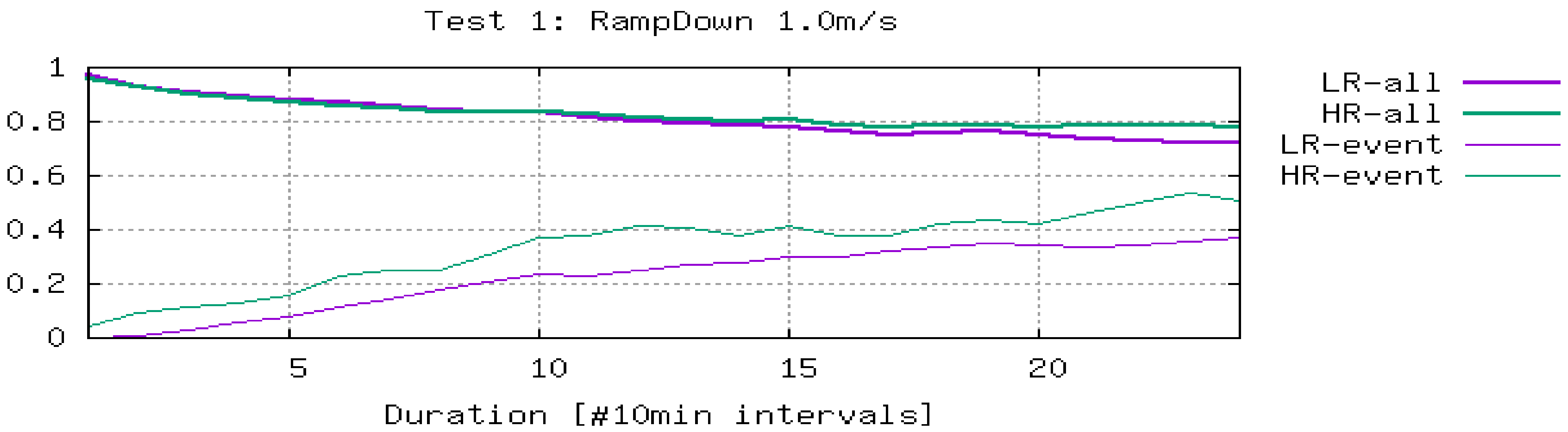

Test 2: RampUp 0.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob



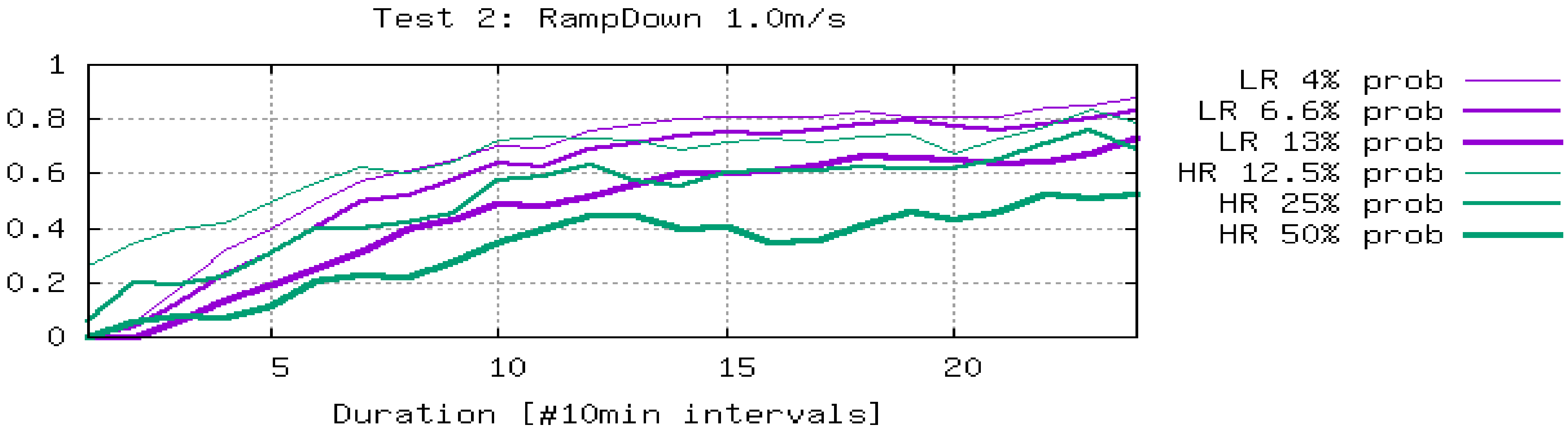
Test 1: RampDown 0.5m/s
LR-all
HR-all
LR-event
HR-event
Duration [#10min intervals]



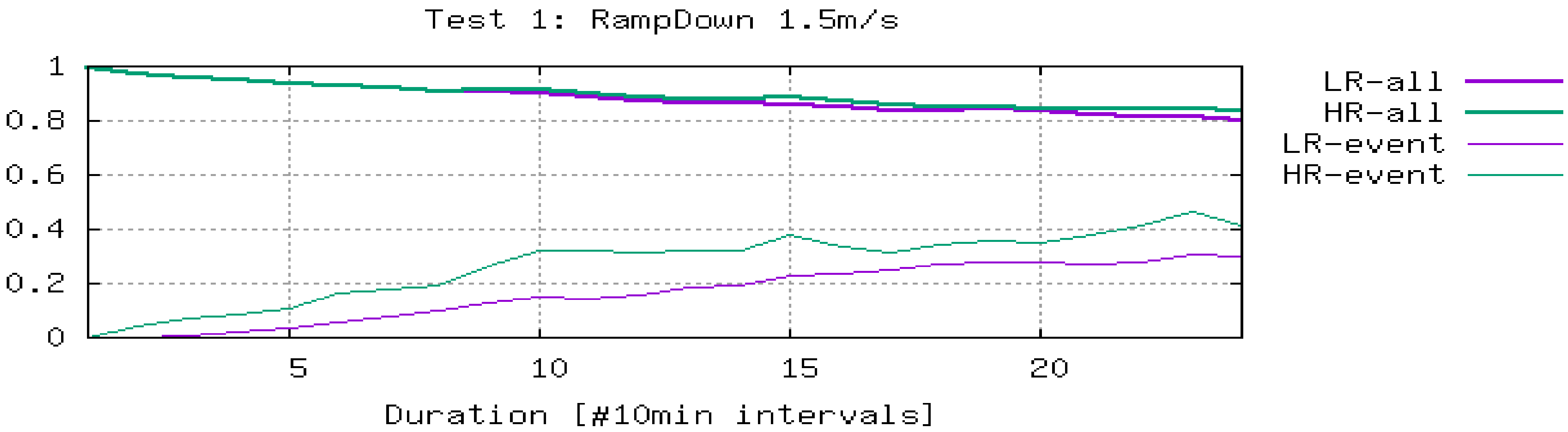
Test 2: RampDown 0.5m/s
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob
Duration [#10min intervals]

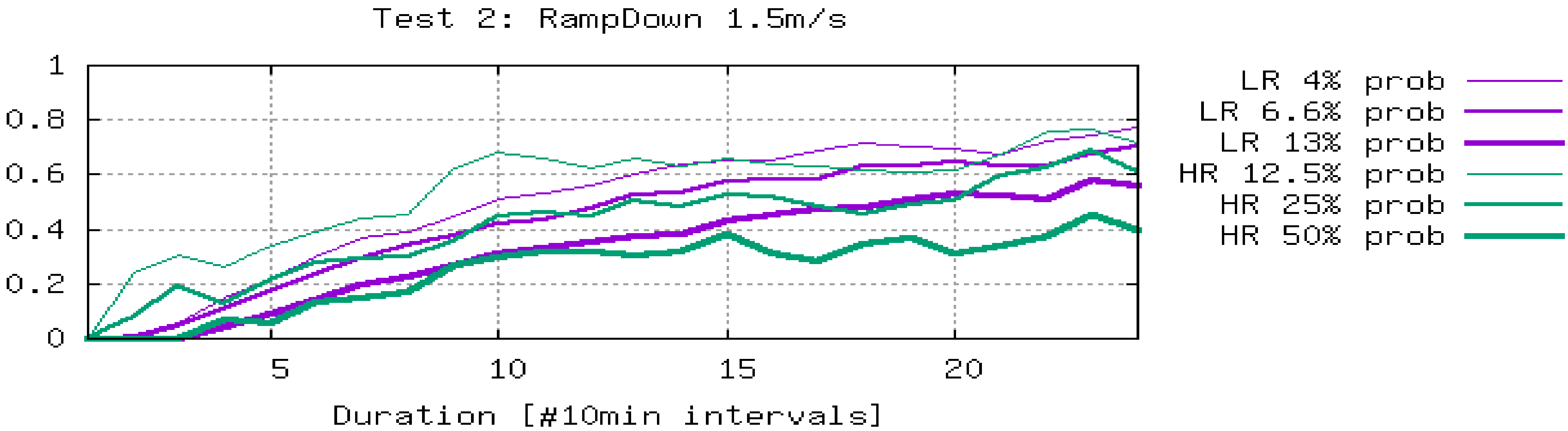


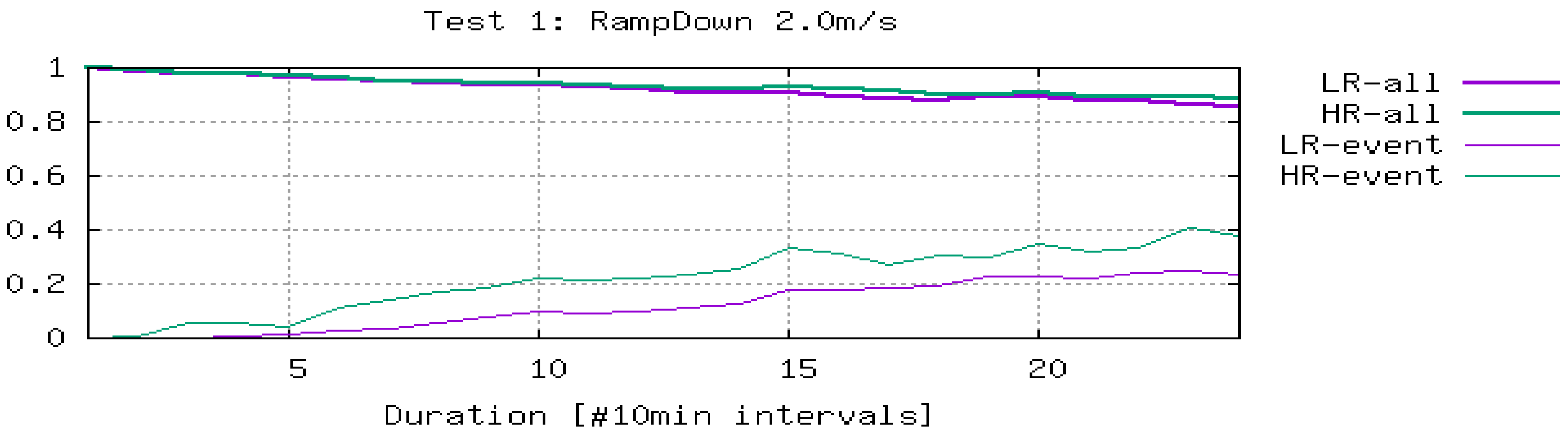
Test 1: RampDown 1.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event

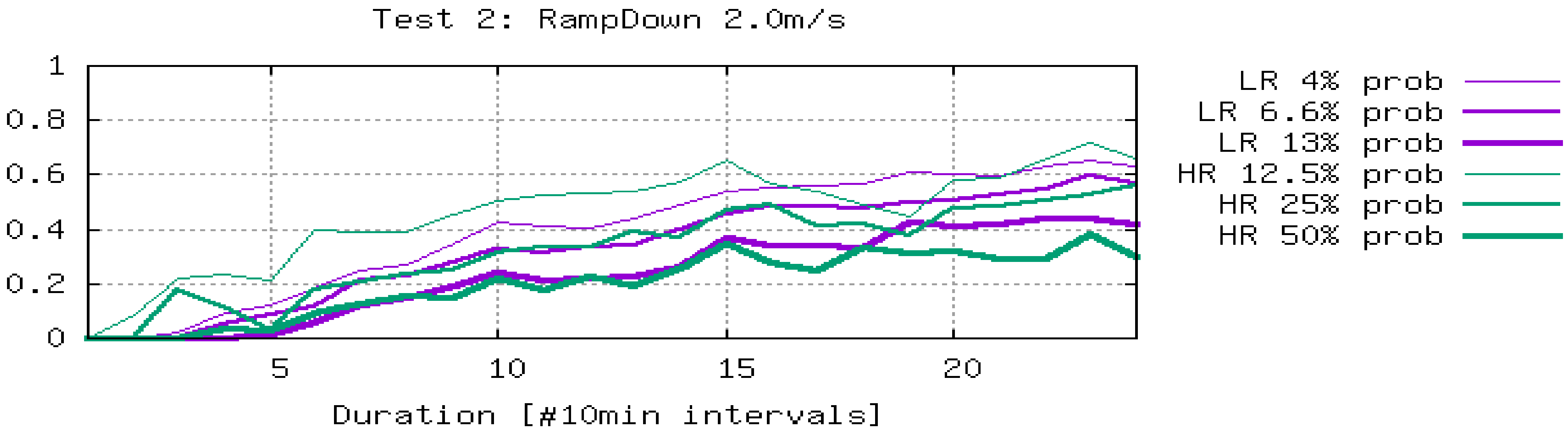


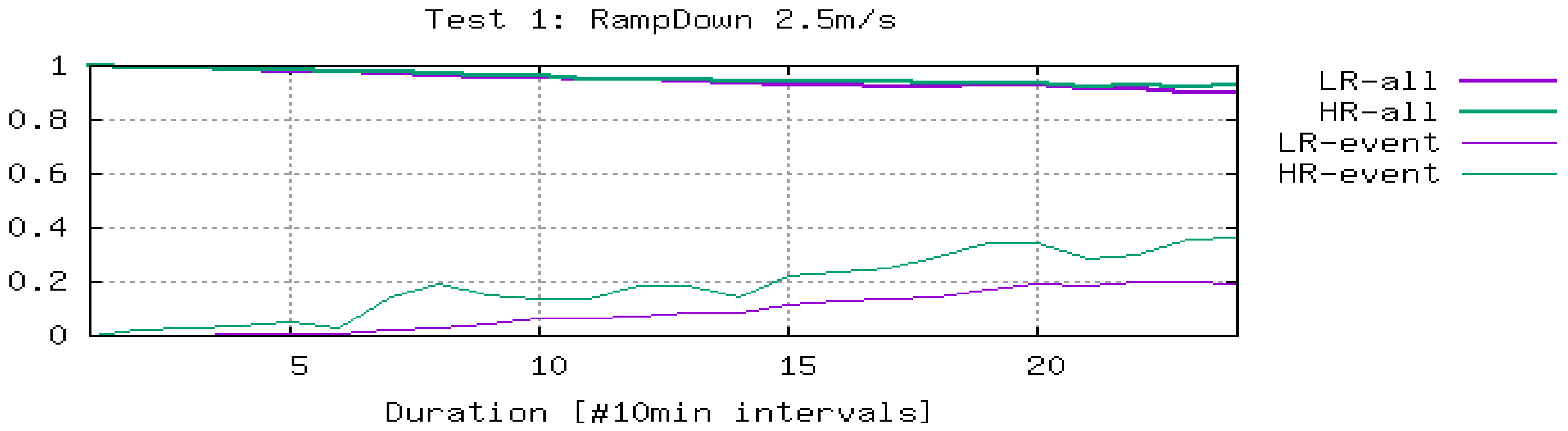
Test 2: RampDown 1.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob

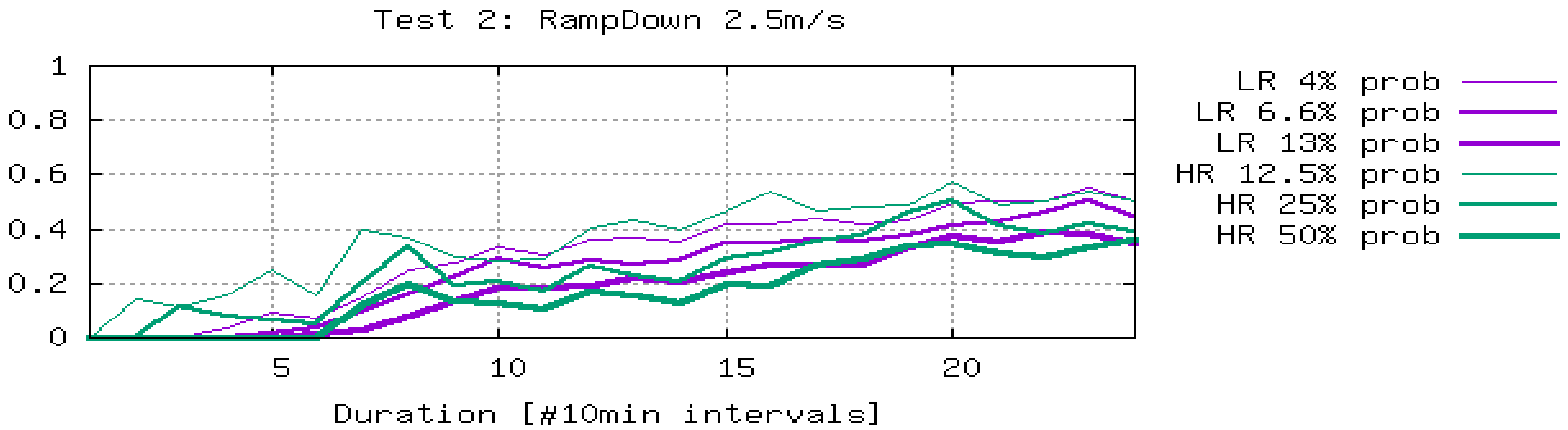

Test 1: RampDown 1.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event


Test 2: RampDown 1.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob


Test 1: RampDown 2.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event


Test 2: RampDown 2.0m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob


Test 1: RampDown 2.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR-all
HR-all
LR-event
HR-event


Test 2: RampDown 2.5m/s
1
0.8
0.6
0.4
0.2
0
5
10
15
20
Duration [#10min intervals]
LR 4% prob
LR 6.6% prob
LR 13% prob
HR 12.5% prob
HR 25% prob
HR 50% prob