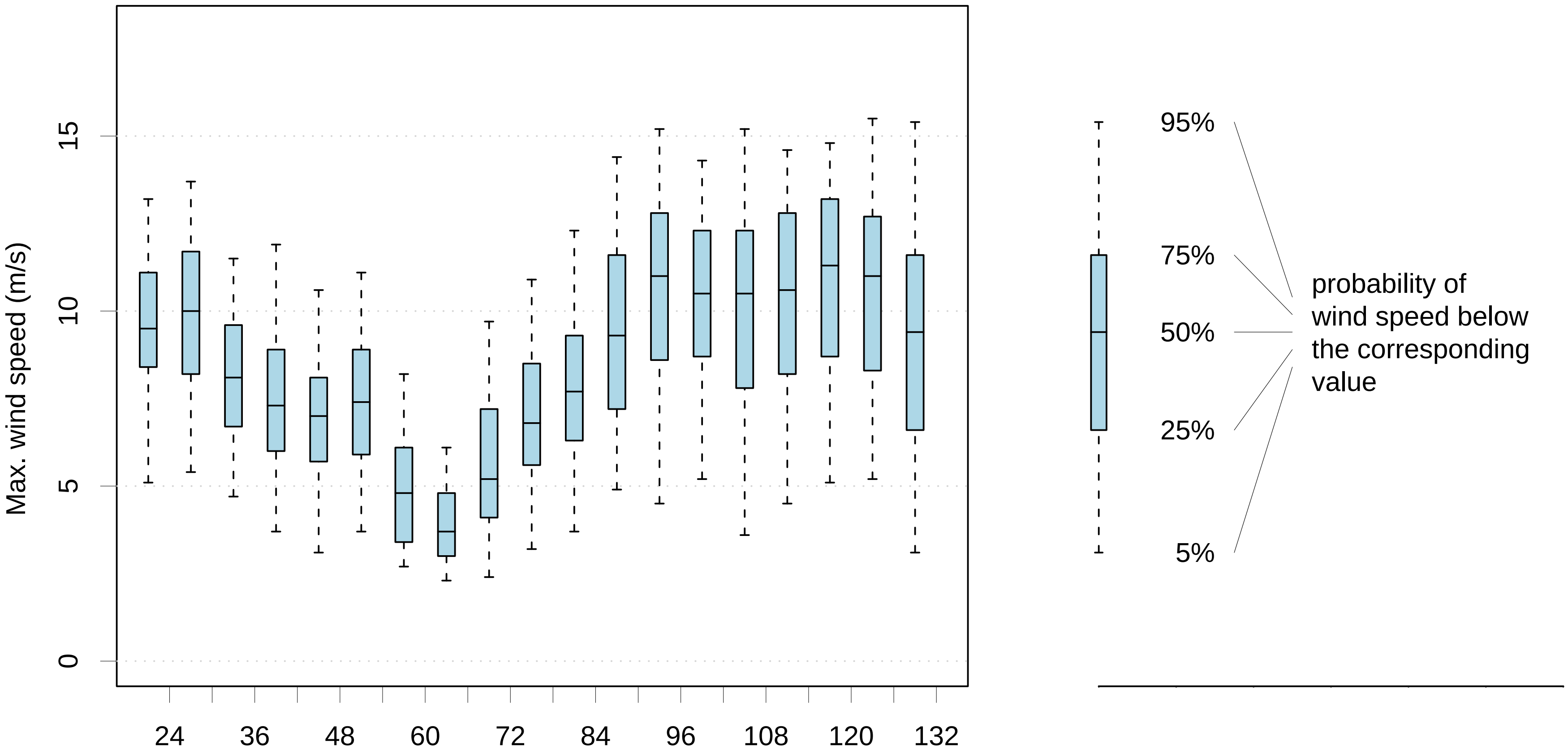
Norwegian Meteorological Institute met.no

Methods for verifying quantile forecasts

John Bjørnar Bremnes

Example



Max. wind speed (m/s)
0
5
10
15
24
36
48
60
72
84
96
108
120
132
95%
75%
50%
25%
5%
probability of
wind speed below
the corresponding
value

Overview

- Reliability
- Sharpness
- Refinement

- Ranking

An example of forecasting wind speed is used throughout

- One station
- 239 cases for the lead time used

Reliability (calibration)

Are the quantile probabilities proper/valid?

Statistic

Fractions of observations below each quantile

Example

	5%	25%	50%	75%	95%
Torungen lighthouse	.083	.227	.541	.762	.923

Hypothesis tests

Does the forecast model produce reliable quantiles?

Each quantile separately

H_0: $p_{true} = p$ (true prob. = quantile prob.)

– binomial test (preferable) or χ^2-test
  - R: binom.test(), chisq.test(), prop.test()

p-values are appropriate for presenting results

– the smaller the p-value, the stronger the evidence that the model is unreliable

All quantiles simultaneously

I. H_0: $p_{true,i} = p_i$ for all quantiles *i*

II. H_0: $p_{true,i} = p_i$ for all intervals *i*
   - The p's are interval probabilities
   - Intervals formed by (between) the quantiles
   - Number of intervals = number of quantiles + 1

- χ^2-tests appropriate for both tests
- 2nd test preferable (no overlapping classes)
- Different p-values!

Example

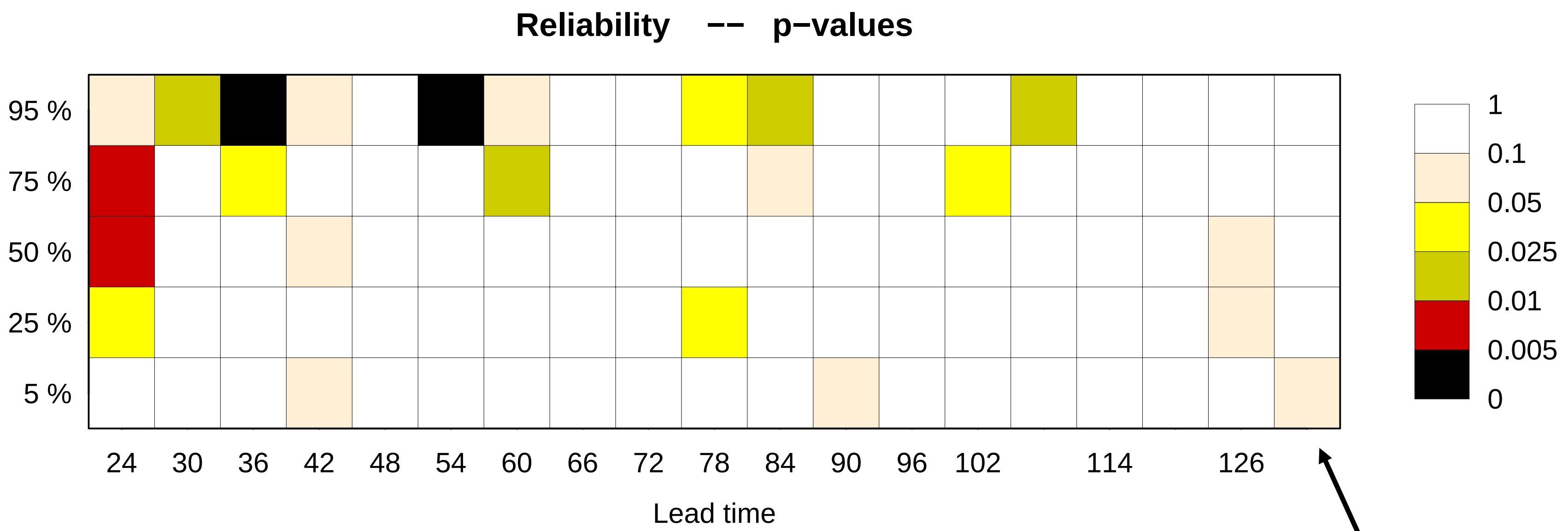
Statistics

	5%	25%	50%	75%	95%
Torungen lighthouse	.083	.227	.541	.762	.923

P-values

	5%	25%	50%	75%	95%	ALL*
Torungen lighthouse	.057	.493	.298	.732	.120	.008

*) p-value based on intervals



The lead time used above

Remarks

- Decision making
  - Choose test(s)
  - Fix significance level
  - Require p-value(s) above this

- χ^2-tests are approximate
  - Thumb rule: expected counts in each cell should be greater than 5 (conservative)

Conditional reliability

- Previous methods only check *overall* reliability (unconditional reliability)
- Quantile probabilities should be valid for *all* forecasts
- Does the reliability depend on
  - Forecasted value?
  - Lead time, time, season, ...?

Stratification of data

- Sort forecasts by e.g. value (for each quantile prob.)
- Group data (e.g. roughly equal sizes)
- Compute statistics for each group

Example

		5%	25%	50%	75%	95%
Quantile value	Low	.117	.333	.550	.767	.950
	Medium	.049	.197	.541	.754	.934
	High	.083	.150	.533	.767	.883

Hypothesis tests

Each quantile separately (by value)

I. H_0: $p_{low} = p_{med} = p_{high}$

II. H_0: $p_{low} = p_{med} = p_{high} = p$ (quantile prob.)

- χ^2-tests can be used in both cases
  - R: prop.test()
- Test II is most complete/relevant
- Same principle for testing all quantiles simultaneously

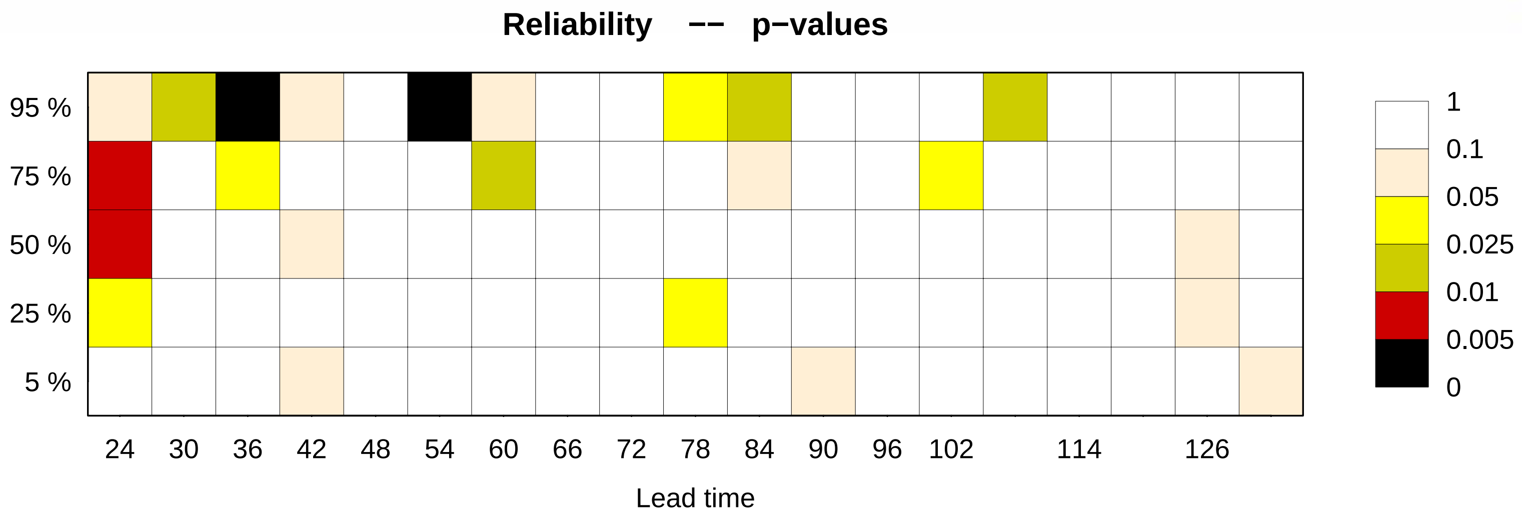
Example

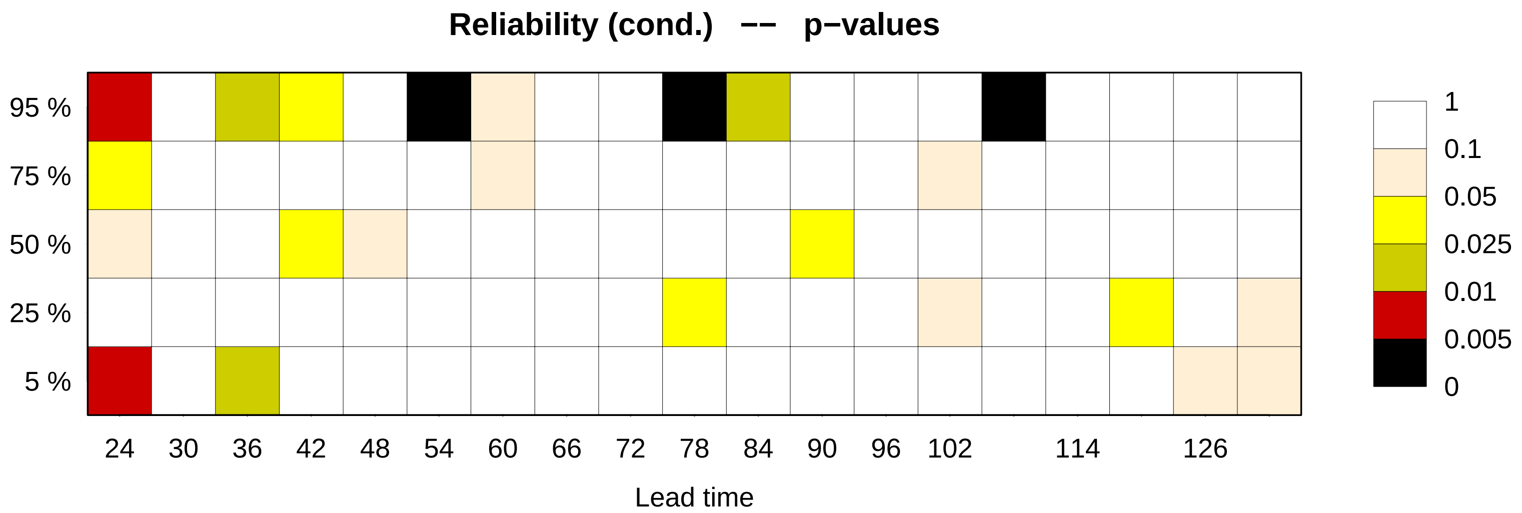
Statistics

		5%	25%	50%	75%	95%
Quantile value	Low	.117	.333	.550	.767	.950
	Medium	.049	.197	.541	.754	.934
	High	.083	.150	.533	.767	.883

P-values

	5%	25%	50%	75%	95%
I (homogenity)	.404	.045	.983	.983	.359
II (joint)	.071	.096	.735	.980	.115
unconditional	.057	.493	.298	.732	.120


Reliability -- p−values
95 %
75 %
50 %
25 %
5 %
24
30
36
42
48
54
60
66
72
78
84
90
96
102
114
126
Lead time
1
0.1
0.05
0.025
0.01
0.005
0


Reliability (cond.) -- p−values
95 %
75 %
50 %
25 %
5 %
24
30
36
42
48
54
60
66
72
78
84
90
96
102
114
126
Lead time
1
0.1
0.05
0.025
0.01
0.005
0

Regression based tests (conditional reliability)

– Logistic regression for each quantile prob.

$$\log\left(\frac{p_{true}}{1-p_{true}}\right) = \alpha_0 + \alpha_1 q_p$$

p-th quantile

H_0: $\alpha_1 = 0$ (no trend)

H_0: $\alpha_1 = 0$ and $\alpha_0 = \log(p/1\text{-}p)$ (no trend and proper prob.)

Likelihood ratio tests?

Sharpness

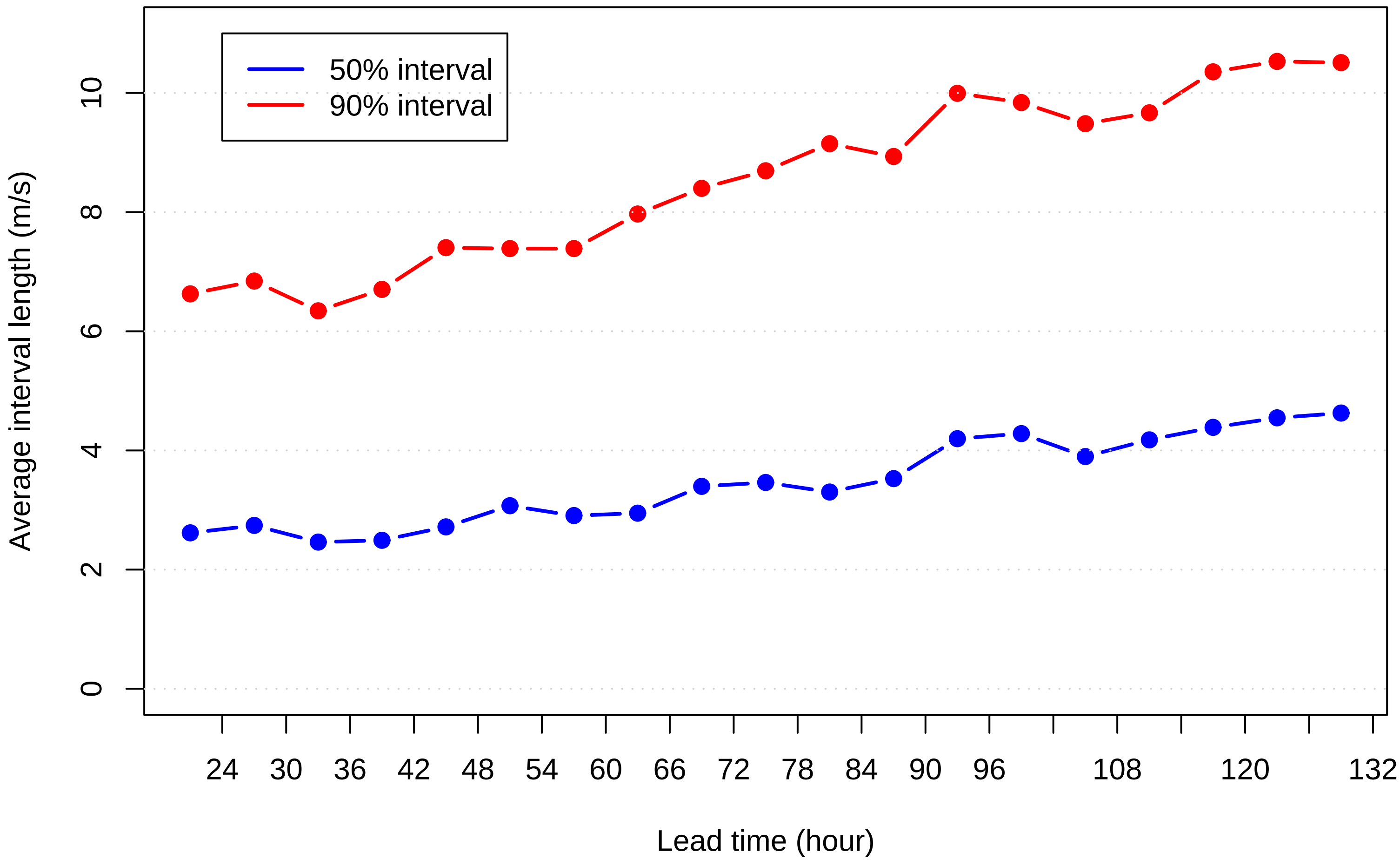
Probability mass should be distributed on short interval(s)

Several quantiles

- Average length(s) of intervals formed by pair(s) of quantiles
    - Ex.: average length of 50% and 90% intervals
    - Bimodality is often penalised too much
    - Empirical distributions provide additional information
    - Single number would be useful for decision making

Single quantile

- Variation as measured by standard deviation or range (as for deterministic forecasts)



50% interval
90% interval
Average interval length (m/s)
0
2
4
6
8
10
24
30
36
42
48
54
60
66
72
78
84
90
96
108
120
132
Lead time (hour)

Refinement / Variation

Information about uncertainty is less important if it is constant

Measures

- Standard deviation (or range) of interval lengths
- Deviation from climate quantiles?

Ranking quantile forecasts

Score functions

- Discrete ranked probability score (RPS) is not suitable
  - Sharpness is not given credit
- Make complete CDFs (and PDFs?) of the quantiles and use CRPS or other scoring rules (not easy)
  - Approximate CRPS by integrating only over the range of the quantiles

Reliability and sharpness

- Require reliability at a given significance level and rank reliable models by average interval length(s)
- Most suitable in the process of making forecast models

Summary

Reliability

- Hypothesis tests useful
- Important to also assess cond. reliability

Sharpness

- Length of forecast intervals

Ranking models important problem

- Scoring rules useful