COST722 WG3
Task 1: Determine how to evaluate the potential of existing methods

Frédéric Atger

April 2004

Task 1: Determine how to evaluate the potential of existing methods

- Goal: an evaluation methodology of statistical techniques
- Expected result: guidance for evaluation
- Resources: available literature
- Completion date: expected November 2004
- Interactions: WG2

- Remark: the proposed methodology can be applied to any forecasting method, statistical or based on numerical modelling

Evaluation or inter-comparison?

- Evaluation often means comparison: a method is *good* compared to a reference, i.e. another method
  - A baseline reference is useful, e.g. persistence
  - But comparing methods is even more meaningful
- Comparison requires a common verification method
- Comparison should involve fog forecasts obtained through numerical modelling (WG2)

Verification principles

- Local verification (not spatial) at met stations
- Event oriented verification
  - Because the issue is not to forecast visibility, but to predict low visibility events when they occur
- Probabilities can be compared to deterministic statements
  - Useful when comparing models and stat methods
- Visibility thresholds (proposal):
  - 200 m (roads)
  - 600 m (airports)
  - 1000 m ("fog")
  - 5000 m ("mist")

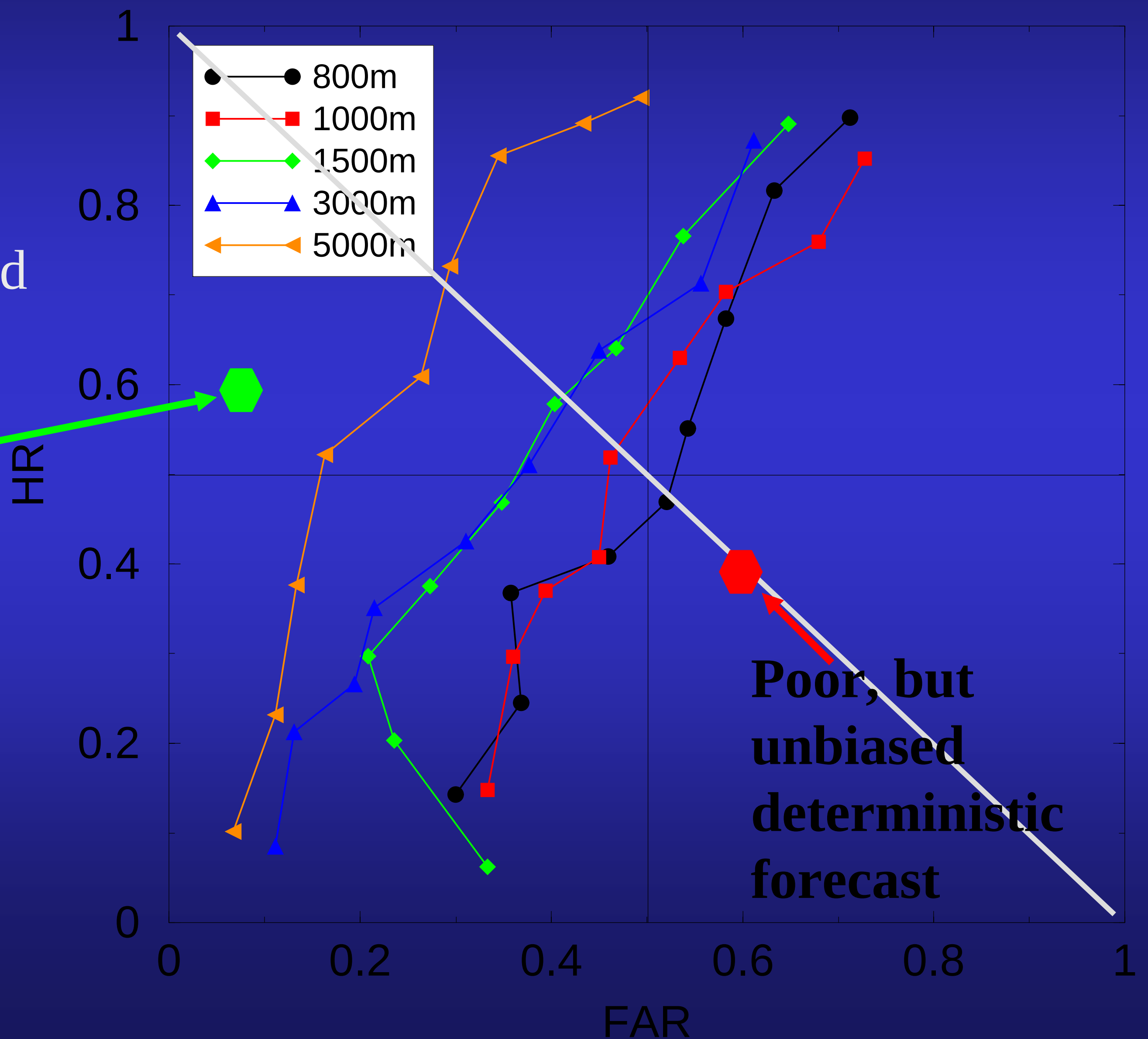
First aspect: detection and false alarms

- A good forecast has a high detection rate
  - Many severe events are successfully forecast
- A good forecast has a low false alarm ratio
  - Few severe event warnings are erroneously issued
- Certain users are very sensitive to false alarms, they would accept a lower detection rate
- Other users would appreciate a higher detection rate even at the expense of many false alarms
- A full view of the performance is given by a 2-dimensional analysis only

ROC or pseudo-ROC ?

- The ROC diagram has many advantages but the meaning of the false alarm rate is not intuitive (proportion of non events that are forecast)
  - Another problem is that false alarm rates are generally very small for rare events (and "good" methods), so that deterministic forecasts appear as a group of points in the bottom left corner
- The pseudo-ROC diagram shows the Hit Rate (HR as in the ROC) vs the False Alarm Ratio (FAR = proportion of warnings that are not justified, i.e. a much more intuitive definition!)
- The pseudo-ROC diagram also shows the bias of deterministic forecasts: there is no bias where HR+FAR=1, i.e. along the diagonal

An example of pseudo-ROC diagram for different visibility thresholds (+18h forecast)



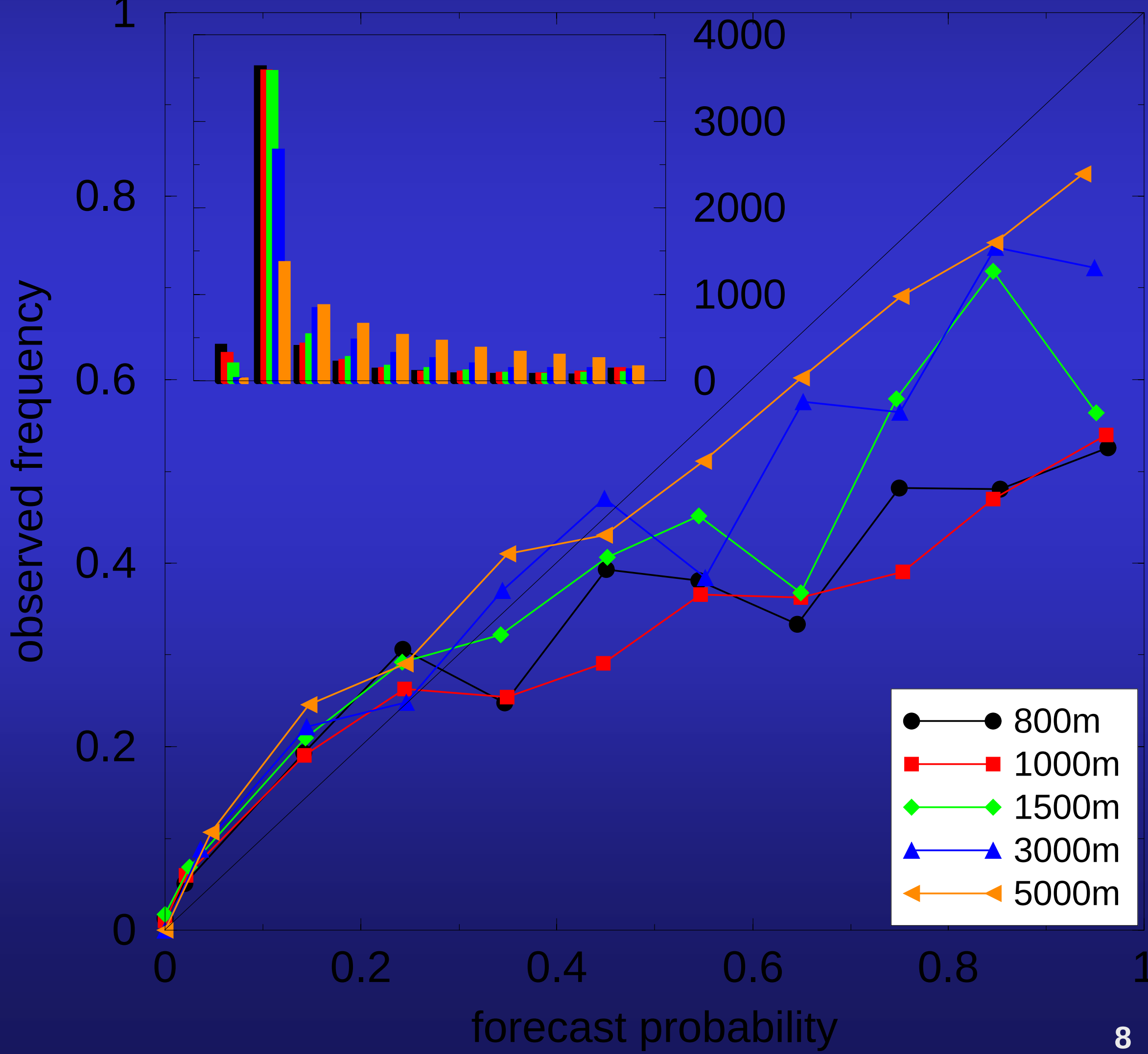
Good, but biased deterministic forecast

Poor, but unbiased deterministic forecast

Second aspect: reliability (of prob forecasts)

- Observed frequencies should reflect forecast probabilities
- Reliability curve

- The distribution of probabilities is useful too: shows the "sharpness" of the forecast, i.e. the propension to "take risks" by forecasting high probabilities of rare events



Sharing the verification software?

- Facilitating results comparison (method, graphics)
- Standard package: ksh scripts and C programs + XMGR (freeware) for the graphics
- Currently run on UNIX and LINUX platforms
- Some work is needed for defining a common format for the input file:
  - ASCII file
  - Line = location, date, observation, probability
  - One file = one lead time / one validity time
  - One file = one threshold
  - Several locations in the same file
- Proposal to be discussed!