ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Verification

Carl Fortelius

HARMONIE training week
SMHI, 19-23 September 2011

Contents

- **Verification - assessing the quality of a forecast**
- **Aspects of quality**
- **Quantitative measures of quality**
  - The HARMONIE verification package
  - Spatial and scale selective methods
- **About the observations**
- **The FMI on line mast monitoring**
- **Links**

Purpose of verification in 1/2

- **Quality assurance**
  - How much can we trust the forecast?
  - How fast are we making progress?
  - Whose forecast is better?

Purpose of verification 2/2

- **Development**
  - Emphasis on structures and processes, i.e. relationships between variables over time
  - Often focusing on systematic errors
  - ”Special” data: profiles, fluxes, physiographic characteristics, etc.

Aspects of quality:

- **Measuring quality is complicated, because a set of forecasts can differ from a set of observations in very many ways**
- **Allan Murphy (1993) distinguished between 9 different attributes of quality:**
  - *Bias, Association, Accuracy, Skill, Reliability, Resolution, Sharpness, Discrimination, Uncertainty*
    For explanations, see: WWRP/WWNE Joint W.G. on Forecast Verification: Forecast Verification: Issues, Methods, and FAQ
    http://www.cawcr.gov.au/projects/verification/

HARMONIE tools and methods 1/6

- **Scatter plots show the correspondence between forecast and observed values** *(association, accuracy, reliability, ...everything)*
- **Histograms show the correspondence between the distributions of forecast and observed values** *(reliability)*
- **Error charts and tables show how some error is distributed in space, and station-wise linear correlation** *(reliability, accuracy, association)*
- **Mean diurnal cycles show how your mean error changes in the course of the day** *(reliability)*
- **Time sequences and vertical profiles show how your data or error characteristic is distributed in time or in the vertical** *(reliability, accuracy)*
- **Error as function of forecast lead time summarises the bias and rms-error and their growth rate over a set of forecasts** *(bias, accuracy)*

HARMONIE tools and methods 2/6

- **A large variety of supplementary scores, based on multi-valued contingency tables transformed into multiple dichotomous contingency tables:**

	observed	
forecast	yes	no
yes	hits (h)	false alarms (fa)
no	misses (m)	correct negatives (cn)

HARMONIE tools and methods 3/6

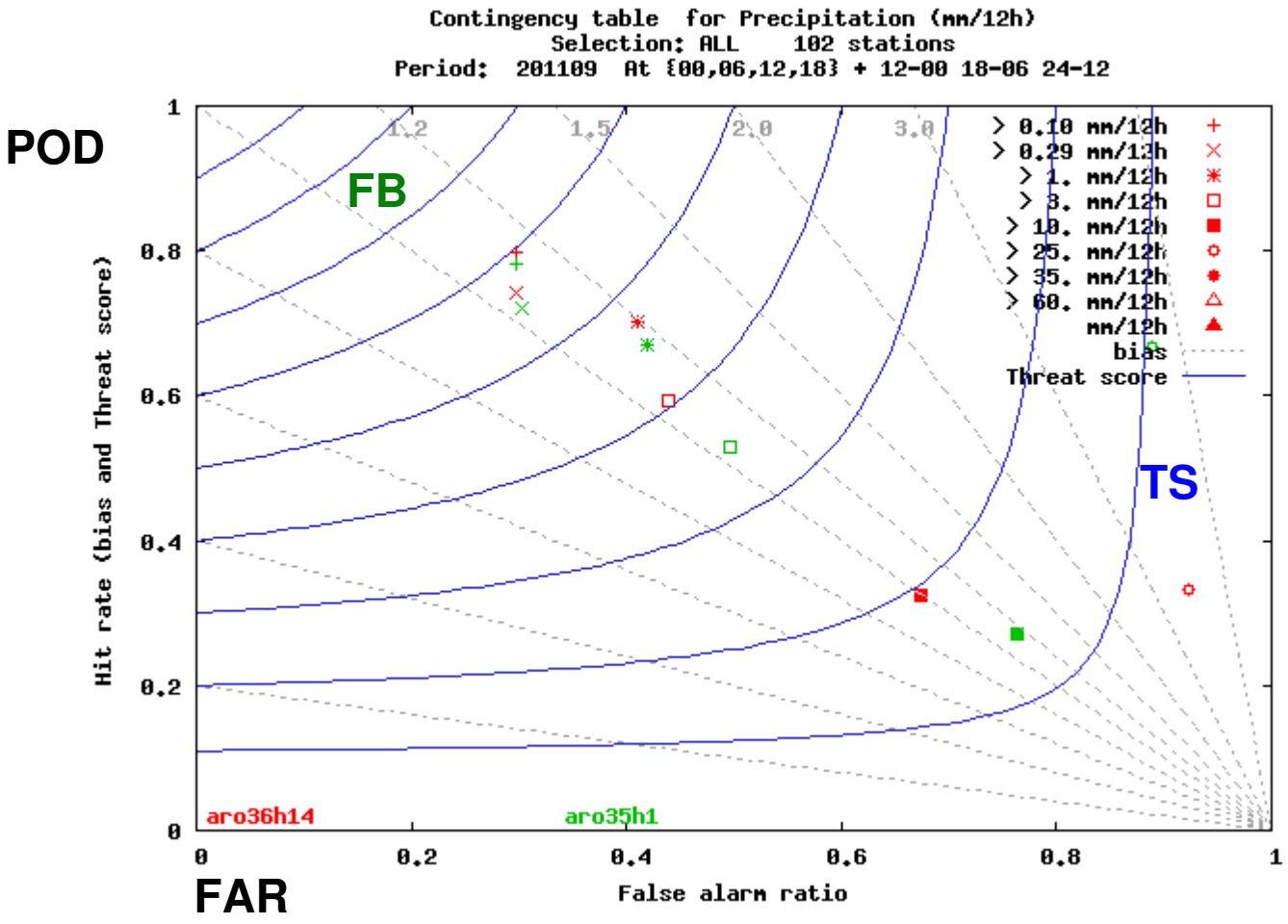
- **Two types of "events" are considered**
  - For the type thresholds, an event is considered to take place when the respective threshold value is exceeded, suited for rainfall
  - For the type classes, an event is considered to take place when the value is between the limits of the respective class, suited for temperature
  - **Note**: narrow classes lead to small populations!

HARMONIE tools and methods 4/6

- Frequency bias (bias score): *(h+fa)/(h+m)*; compares the frequency of predicted events to the frequency of observed events
- Hit rate (probability of detection): *h/(h+m)*; What fraction of the observed events were correctly forecast
- False alarm ratio: *fa/(h+fa)*; What fraction of the predicted events did not occur
- Threat score (critical success index): *h/(h+m+fa)*;How well did the predicted events correspond to the observed events
- The above scores are combined into so-called ”Wilson diagrams” (After Clive Wilson)

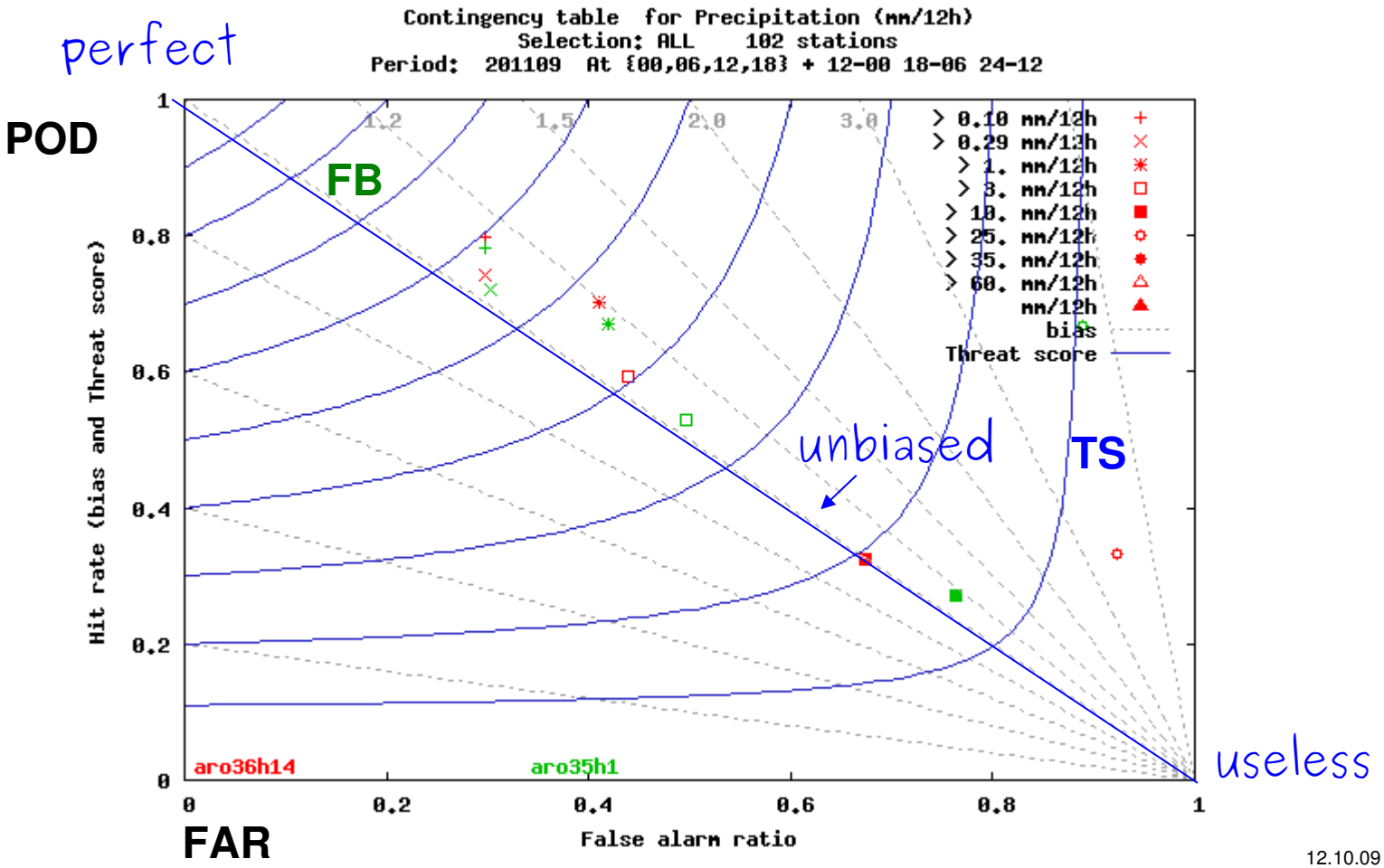
HARMONIE tools and methods 5/6

"Wilson-diagram" for FMI HARMONIE parallel test:



HARMONIE tools and methods 5/6

"Wilson-diagram" for FMI HARMONIE parallel test:



HARMONIE tools and methods 6/6

- Rare events tend to score badly in the Wilson-diagram. The Equitable threat score takes into account the number of random hits *(R)* and is less sensitive to climatology: ETS=*(h-R)/(h+m+fa-R), R=(h+m)(h+fa)*/(h+m+fa+cn). Often used in verification of precipitation
- Hansen-Kuipers score: *(h/(h+m) - fa/(fa+cn))*, How well did the forecast separate events from non-events
- More scores are available, and can be added into the script: harmonie-36h1.4/util/monitor/scr/contingency2gnuplot.pl

Spatial and scale selective methods 1/3

- **Measures based on point-by-point intercomparison are blind to many attributes of a useful forecast (list by Beth Ebert, 2006)**
  - Resembles the observations on the broader scale
  - Predicts an event somewhere near where it was observed
  - Predicts the event over the same area (i.e., with the same frequency) as observed
  - Has a similar distribution of intensities as the observations
  - Looks like what a forecaster would have predicted if she'd had knowledge of the observations

Spatial and scale selective methods 2/3

- Non-local methods for the verification of quantitative precipitation forecats esp. on the meso-scale have been developed
  - Neighbourhood methods find the scale where the forecast starts to have skill
  - Scale-separation methods compare skill at different scales
- The SRNWP-V programme carried out an inventory and made recommendations for their use.(*SRNWP-V D3: Inventory and recommendations of "new" scale selective verification methods, available from the EUMETNET portal)*

Spatial and scale selective methods 3/3

Method	References	Scores	Decision model for useful forecast
Upscaling	Zepeda-Arce et al.2000; Weygandt et al. 2004;Yates et al., 2006	Bias, Threat, ETS	Forecast resembles observations when both averaged to larger scales
Minimum coverage	Damrath, 2004	POD, FAR, ETS	Predicts event over a minimum fraction of region
Fuzzy logic	Damrath, 2004	POD, FAR, ETS	More correct than incorrect
Joint probability	Ebert, 2002	POD, FAR, ETS	More correct than incorrect
Multi event contingency table	Atger, 2001	ROC, Pierce	Predicts at least one event close to observed
Structure, Amplitude, Location (SAL)	Wernli et al., 2008	Struct. Amp. Loc.	structural similarity, total areal amount centre of gravity

available in HARMONIE

→ : recommended by SRNWP-V

Spatial and scale selective methods 3b/3

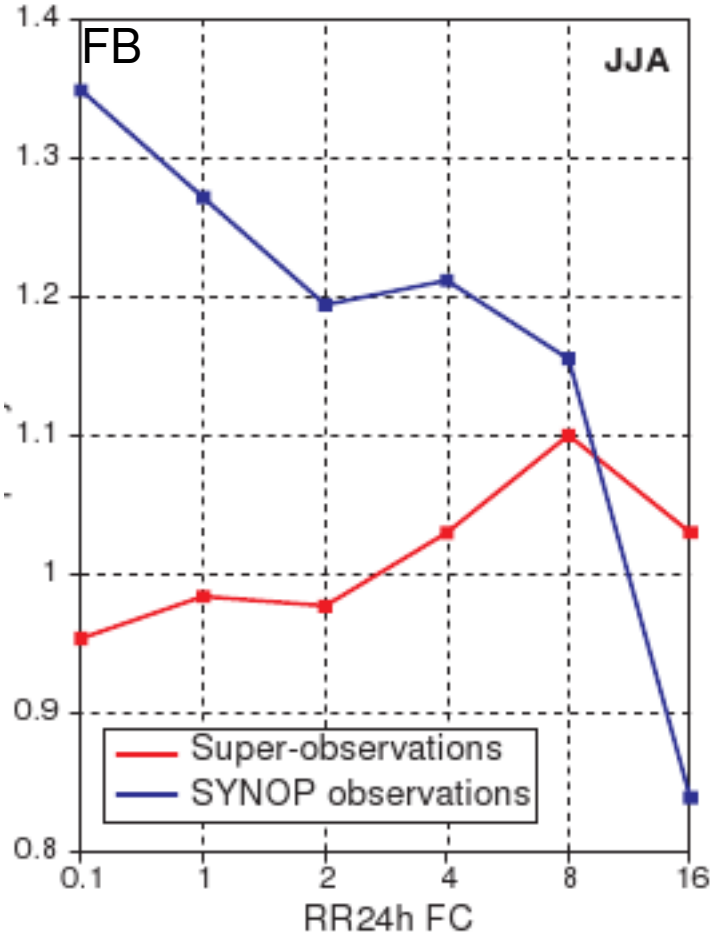
Intensity-scale	Casati et al, 2004	Skill score	Forecast structure lower error than random arrangement of observations
Fractional skill score	Roberts and Lean, 2008	FSS	Over region similar frequency observed and forecast
Pragmatic	Theis et al, 2005	Brier, Brier Skill (BS, BSS)	Useful forecasts has high probability of detecting events and non-events
Practically perfect hindcast	Brooks et al, 1998	Threat scores, ETS ratio	Resembles forecast that would have been issued by forecaster with perfect knowledge of observations befoerhand
Conditional square root of RPS	Germann and Zawadzki, 2004	CSRR	High probability of matching observed value
Area-related RMSE	Rezacova et al, 2007	RMSE	Similar intensity distribution as observed.

Observations vs. forecast

- **Spatial scale**
  - local vs. grid box average
- **Height above ground**
  - e.g. light houses often measure “10 m wind” at 20-50 m
- **Scenery**
  - Is it right to compare local reports to grid averages, or should representative tiles be used, e.g for screen temperature?
  - what should be issued as a forecast?

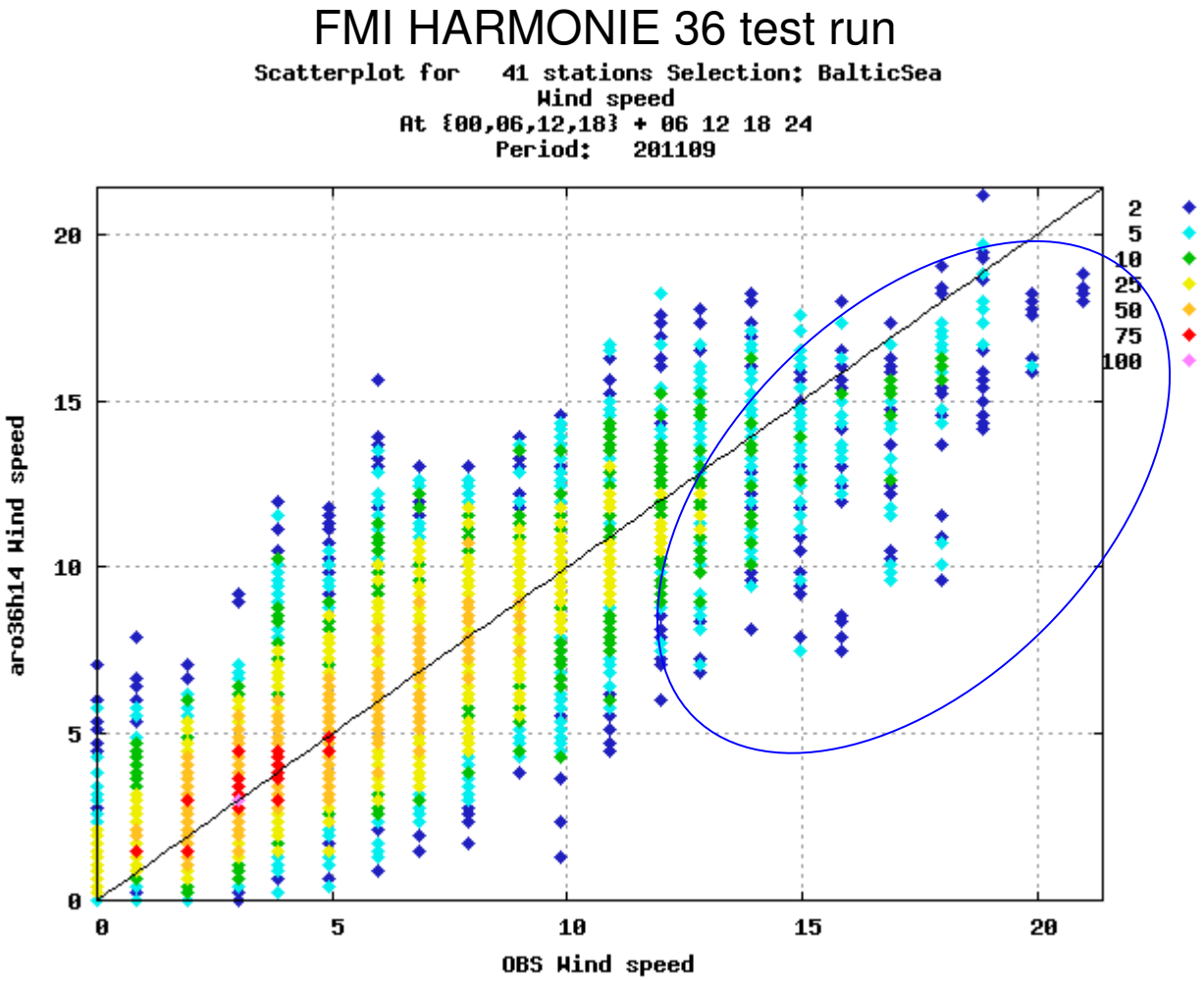
Spatial scale

- Models over-predict weak precipitation and under-predict high intensities?
- Figure from: *Anna Ghelli and François Lalaurette: Verifying precipitation forecasts using upscaled observations, ECMWF Newsletter Number 87 – Spring 2000*



France, JJA 1997

Measuring height



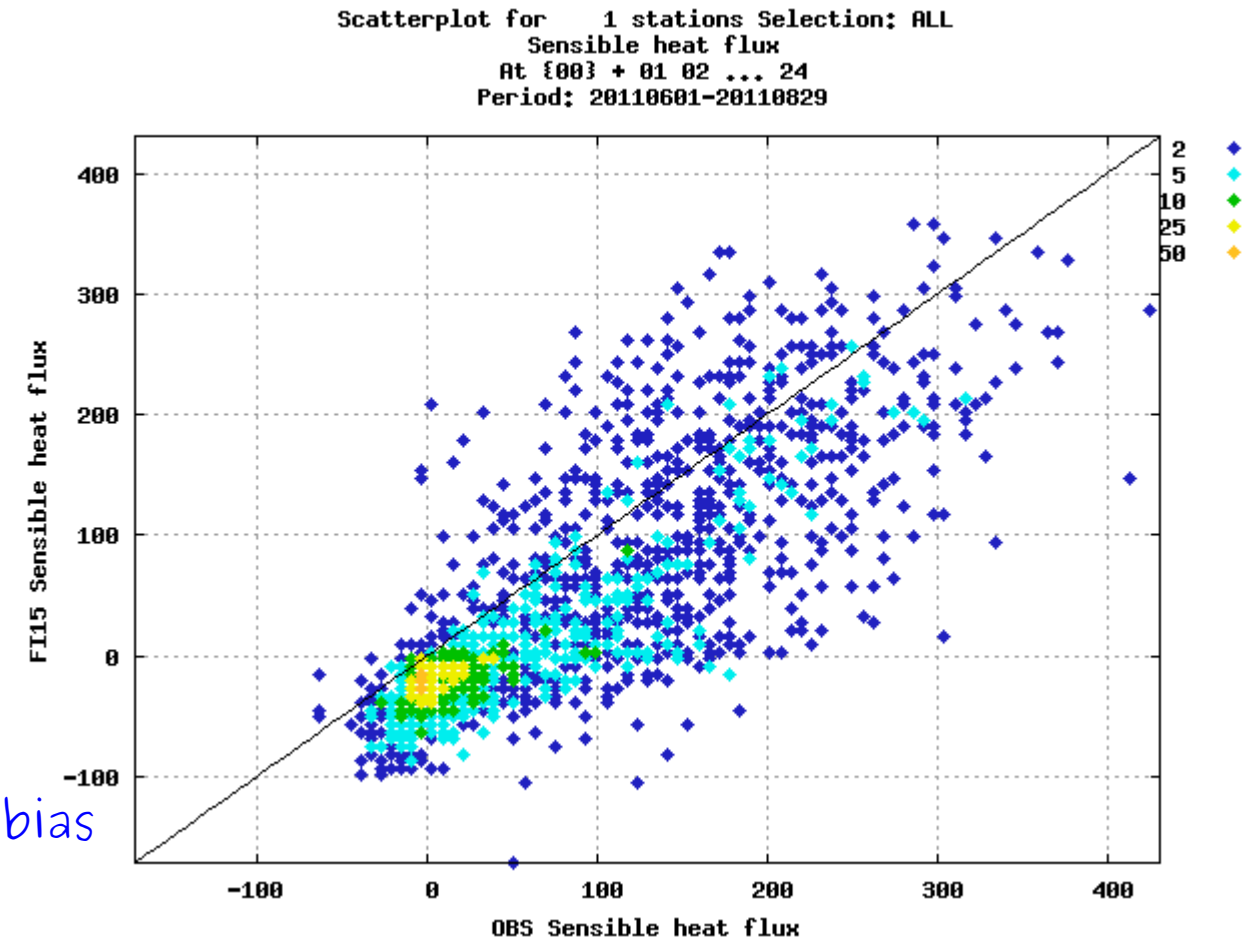
Model error or the effect of high towers?

Diagnostic verification 1

- The FMI Mast monitoring compares near surface weather elements and fluxes from several NWP systems at several observatories in the form of:
- daily real time plots
- seasonal summaries (a newly added feature)

Diagnostic verification 2:

The sensible heat flux at **Sodankylä** in summer 2011: HIRLAM RCR vs measured

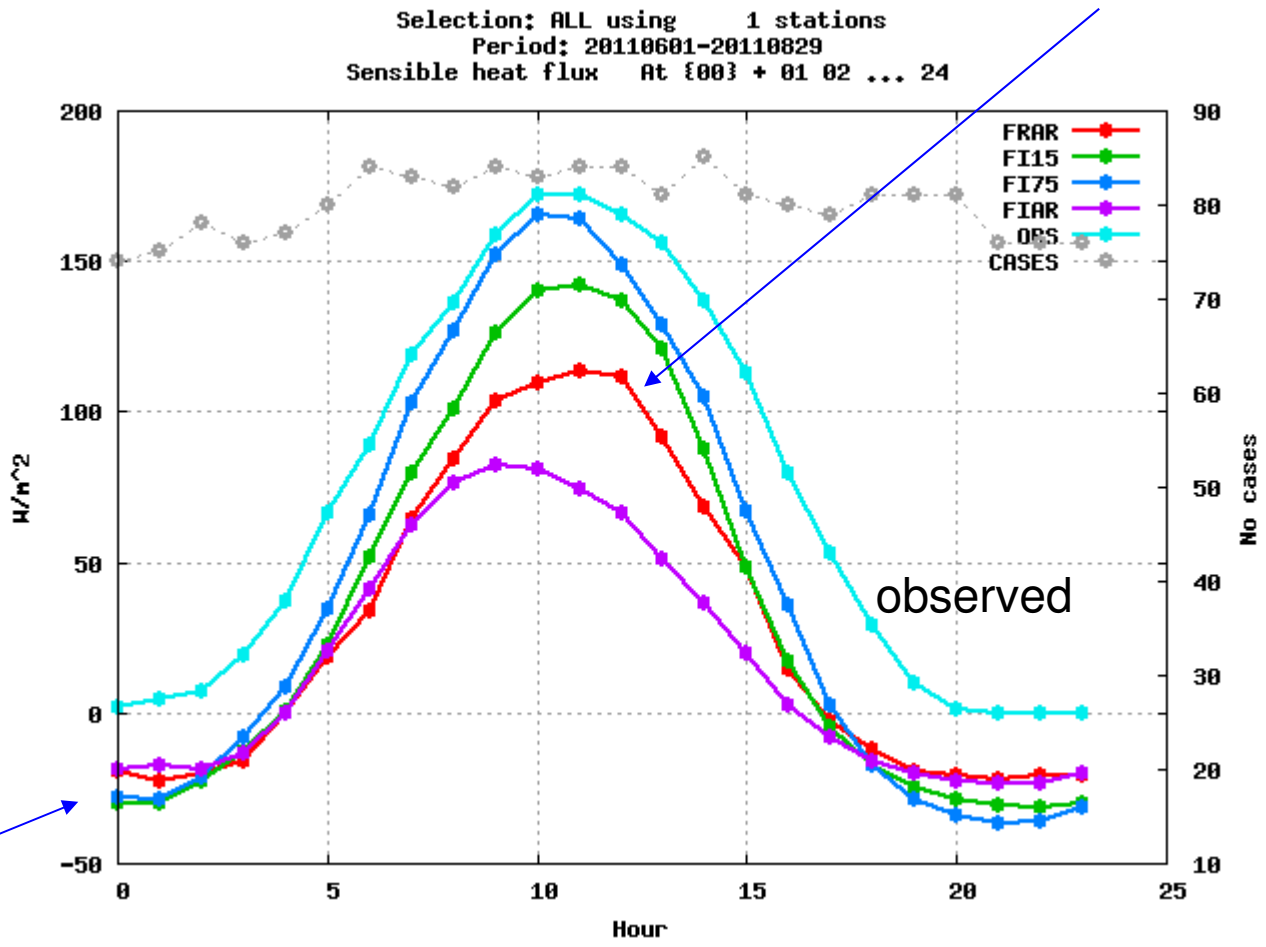


Negative bias

Diagnostic verification 3:

All models underestimate the day time flux

The sensible heat flux at **Sodankylä** in summer 2011: Mean diurnal cycles



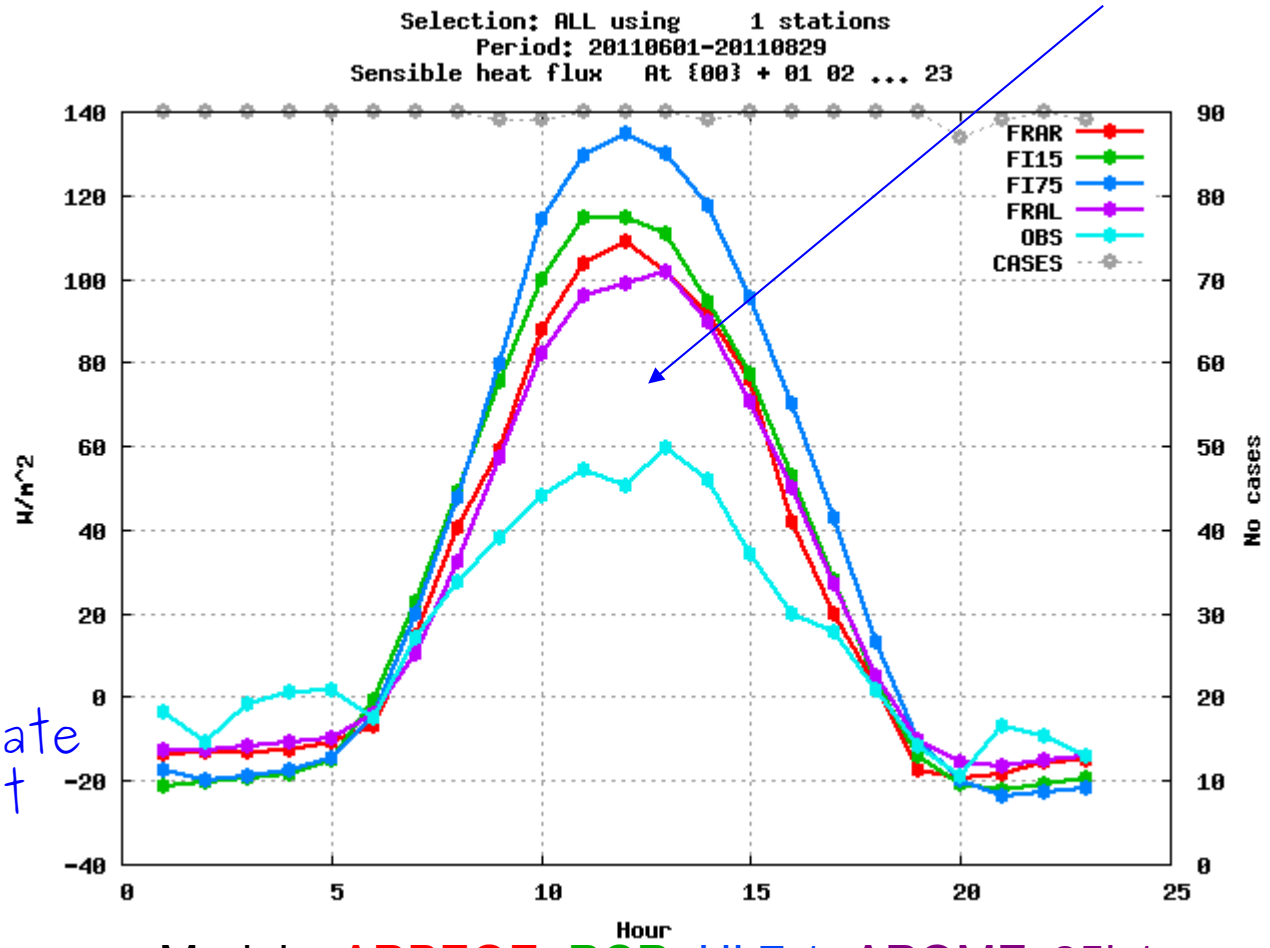
All models give spurious downward flux at night

Models: ARPEGE, RCR, HL7.1, AROMEc35h1

Diagnostic verification 4:

All models overestimate the day time flux

The sensible heat flux at **Cabauw** in summer 2011: Mean diurnal cycles



All models overestimate the downward flux at night

Models: ARPEGE, RCR, HL7.1, AROMEc35h1

Summary

- **Complexity:** Good verification involves looking at the forecasts and the data from many sides. Summary measures are handy but may be difficult to interpret. Stratifying your data facilitates interpretation but reduces sample sizes
- **Spatial methods:** point by point intercomparison is blind to many desirable attributes of a forecast
- **The truth:** Do your data and forecast represent the same thing?
- **Diagnostic verification** can point to model errors

Links

- The Centre for Australian Weather and climate research, Forecast Verification: Issues, Methods and FAQ:http://www.cawcr.gov.au/projects/verification/
- WMO WWRP Forecast verification research: http://www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html
- talk of E Ebert on spatial methods: http://www.mmm.ucar.edu/events/qpf06/QPF/Session6/ebert_FuzzyForecastVerification.pdf
- The verification of ECMWF forecasts: http://www.ecmwf.int/products/forecasts/guide/The_verification_of_ECMWF_forecasts.html