Introduction to Forecast Verification

Tressa L. Fowler, Tara L .Jensen, Barbara G. Brown

National Center for Atmospheric Research
Boulder Colorado USA

Developmental Testbed Center

Outline

- **Basic verification concepts**
  - **What is verification?**
  - **Why verify?**
  - **Identifying verification goals**
  - **Forecast "goodness"**
  - **Designing a verification study**
  - **Types of forecasts and observations**
  - **Matching forecasts and observations**
  - **Verification attributes**
  - **Miscellaneous issues**
  - **Questions to ponder: Who? What? When? Where? Which? Why?**
- **Categorical verification statistics**
  - **Contingency tables**
  - **Thresholds**
  - **Skill scores**
  - **Receiver Operating Characteristic (ROC) curves**
- **Continuous verification statistics**
  - **Joint and Marginal distributions**
  - **Scatter plots**
  - **Discrimination plots**
  - **Conditional statistics and plots**
  - **Commonly used verification statistics**

What is verification?

- Verification is the process of comparing forecasts to relevant observations
  - Verification is one aspect of measuring forecast ***goodness***
- Verification measures the ***quality*** of forecasts (as opposed to their ***value***)
- For many purposes a more appropriate term is "***evaluation***"

Why verify?

- Purposes of verification (traditional definition)

  - Administrative purpose
    - Monitoring performance
    - Choice of model or model configuration (has the model improved?)

  - Scientific purpose
    - Identifying and correcting model flaws
    - Forecast improvement

  - Economic purpose
    - Improved decision making
    - "Feeding" decision models or decision support systems

Identifying verification goals

What *questions* do we want to answer?

- Examples:
  - ✓ In what locations does the model have the best performance?
  - ✓ Are there regimes in which the forecasts are better or worse?
  - ✓ Is the probability forecast well calibrated (i.e., reliable)?
  - ✓ Do the forecasts correctly capture the natural variability of the weather?

Other examples?

Identifying verification goals (cont.)

- What forecast performance *attribute* should be measured?
  - Related to the *question* as well as the type of forecast and observation
- Choices of verification statistics, measures, graphics
  - Should match the type of forecast and the attribute of interest
  - Should measure the quantity of interest (i.e., the quantity represented in the question)

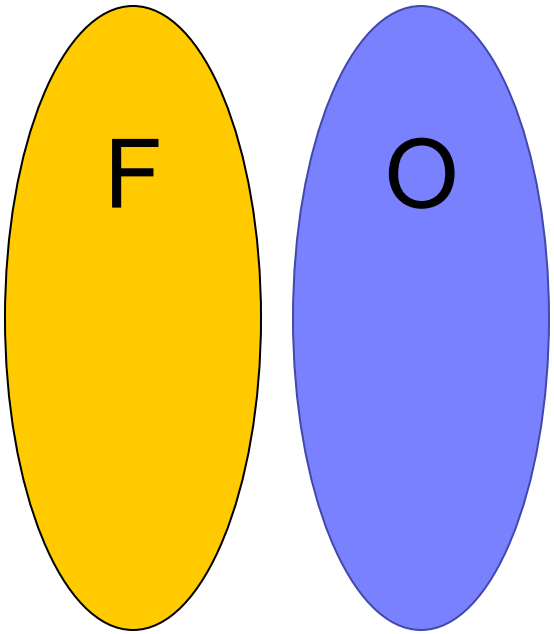
Forecast "goodness"

- Depends on the quality of the forecast

AND

- The user and his/her application of the forecast information

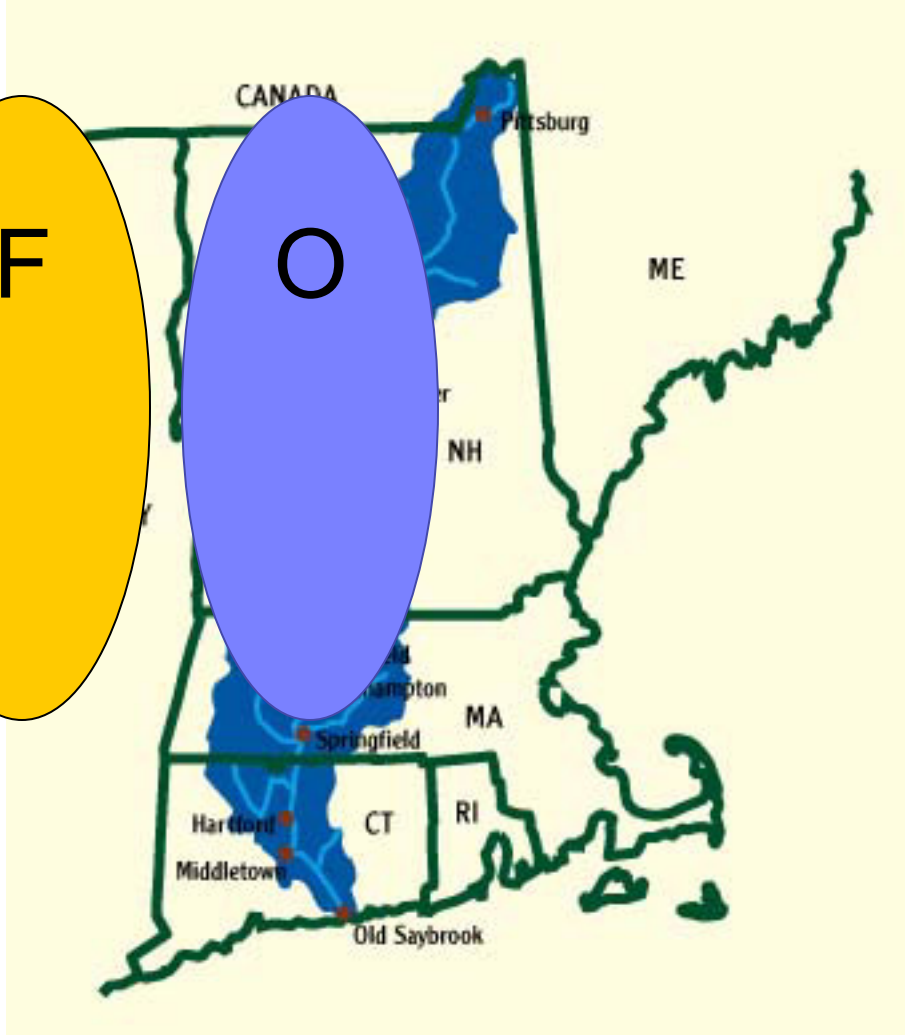
Good forecast or bad forecast?



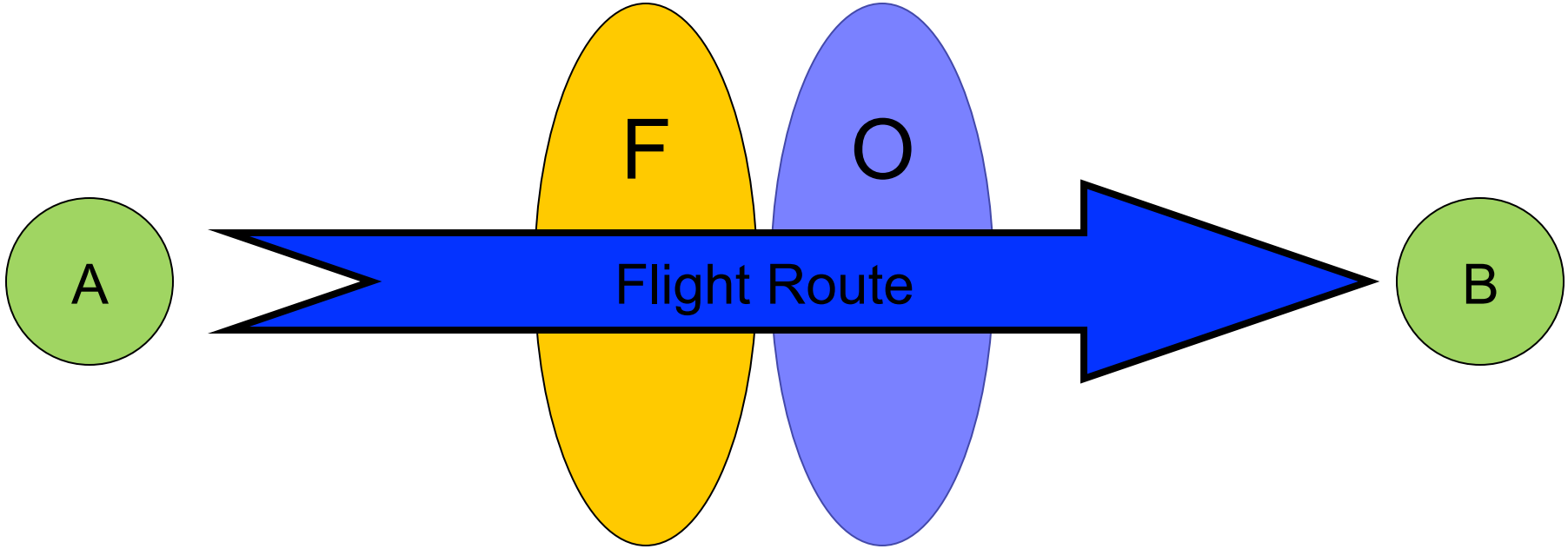
Many verification approaches would say that this forecast has NO skill and is very inaccurate.

Good forecast or Bad forecast?

If I'm a water manager for this watershed, it's a pretty bad forecast…



Good forecast or Bad forecast?



If I'm an aviation traffic strategic planner…

It might be a pretty good forecast

Different users have different ideas about what makes a forecast good

Different verification approaches can measure different types of "goodness"

Forecast "goodness"

- Forecast quality is only one aspect of forecast "goodness"
- Forecast value is related to forecast quality through complex, non-linear relationships
  - In some cases, *improvements in forecast quality (according to certain measures) may result in a degradation in forecast value for some users!*
- ***However*** - Some approaches to measuring forecast quality can help understand goodness
  - Examples
    - ✓ Diagnostic verification approaches
    - ✓ New features-based approaches
    - ✓ Use of multiple measures to represent more than one attribute of forecast performance
    - ✓ Examination of multiple thresholds

Basic guide for developing verification studies

Consider the users…

- … of the forecasts
- … of the verification information

- What aspects of forecast quality are of interest for the user?
  - Typically (always?) need to consider multiple aspects

Develop verification questions to evaluate those aspects/attributes

- *Exercise*: What verification questions and attributes would be of interest to …
  - … operators of an electric utility?
  - … a city emergency manager?
  - … a mesoscale model developer?
  - … aviation planners?

Basic guide for developing verification studies

Identify *observations* that represent the ***event*** being forecast, including the

- Element (e.g., temperature, precipitation)
- Temporal resolution
- Spatial resolution and representation
- Thresholds, categories, etc.

Observations are not truth

- We can't know the complete "truth".
- Observations generally are more "true" than a model analysis (at least they are relatively more independent)
- Observational uncertainty should be taken into account in whatever way possible
  - ✓In other words, how well do adjacent observations match each other?

Observations might be garbage if

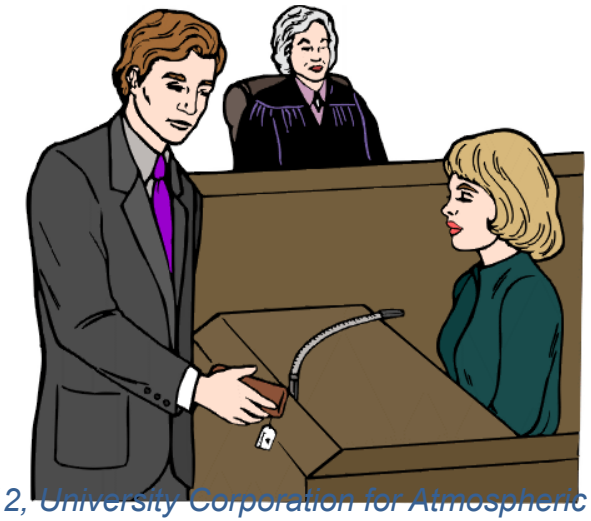
- Not Independent (of forecast or each other)
- Biased
  - Space
  - Time
  - Instrument
  - Sampling
  - Reporting
- Measurement errors
- Not enough of them

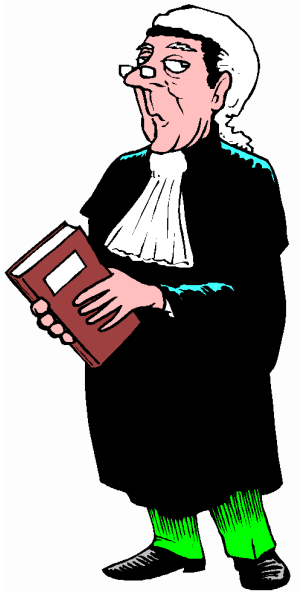
Basic guide for developing verification studies

Identify multiple *verification attributes* that can provide answers to the questions of interest

Select *measures and graphics* that appropriately measure and represent the attributes of interest

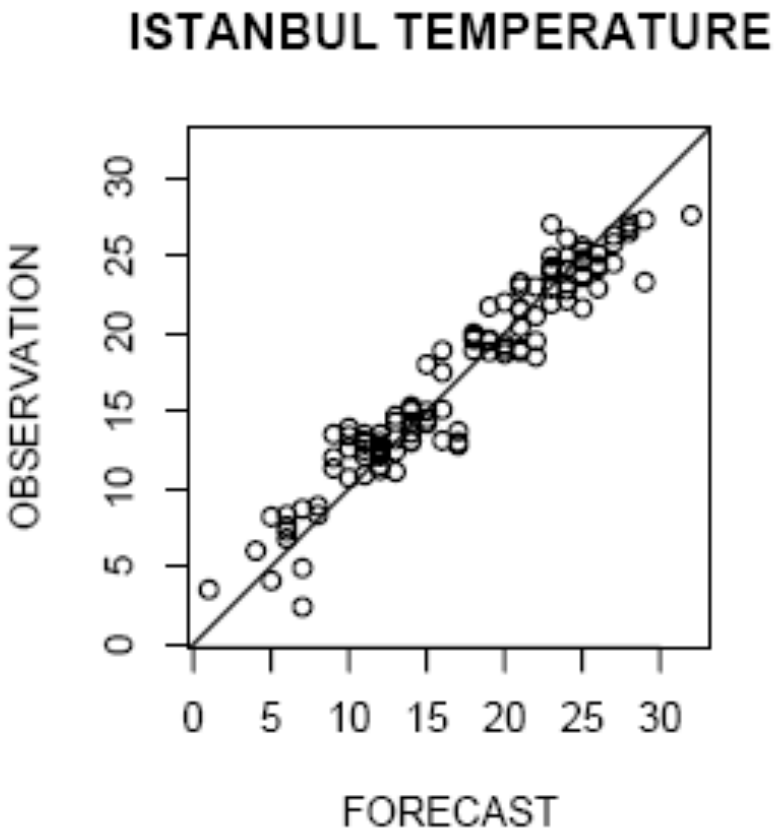
Identify a *standard of comparison* that provides a reference level of skill (e.g., persistence, climatology, old model)





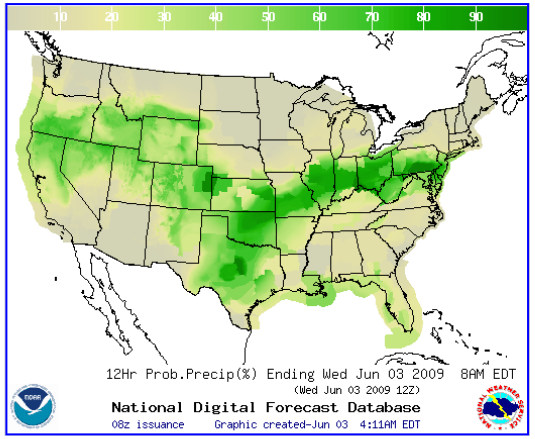
Types of forecasts, observations

- Continuous
  - Temperature
  - Rainfall amount
  - 500 mb height
- Categorical
  - Dichotomous
    - ✓ Rain vs. no rain
    - ✓ Strong winds vs. no strong wind
    - ✓ Night frost vs. no frost
    - ✓ Often formulated as Yes/No
  - Multi-category
    - ✓ Cloud amount category
    - ✓ Precipitation type
  - May result from *subsetting* continuous variables into categories
    - ✓ *Ex: Temperature categories of 0-10, 11-20, 21-30, etc.*



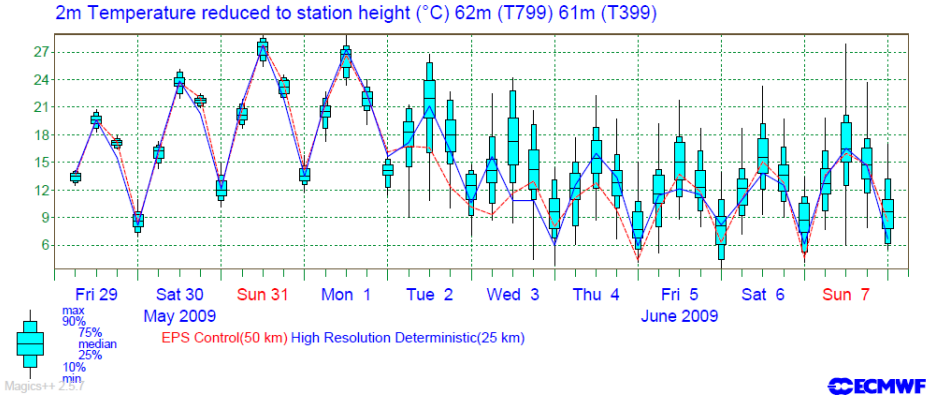
Types of forecasts, observations

- Probabilistic
  - Observation can be dichotomous, multi-category, or continuous
    - Precipitation occurrence – Dichotomous (Yes/No)
    - Precipitation type – Multi-category
    - Temperature distribution - Continuous
  - Forecast can be
    - Single probability value (for dichotomous events)
    - Multiple probabilities (discrete probability distribution for multiple categories)
    - Continuous distribution
  - For dichotomous or multiple categories, probability values may be limited to certain values (e.g., multiples of 0.1)



2-category precipitation forecast (PoP) for US

- Ensemble
  - Multiple iterations of a continuous or categorical forecast
    - May be transformed into a probability distribution
  - Observations may be continuous, dichotomous or multi-category



ECMWF 2-m temperature meteogram for Helsinki

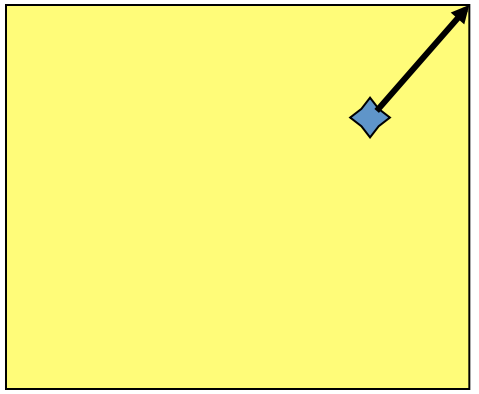
Matching forecasts and observations

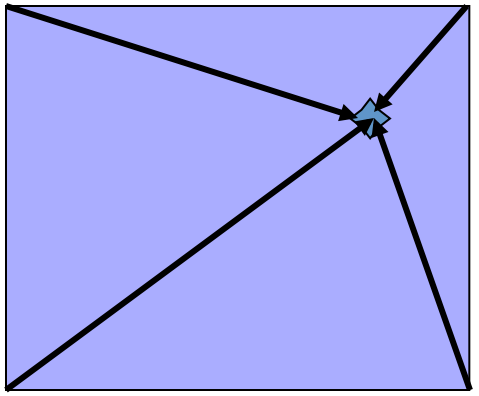
- May be the *most difficult* part of the verification process!
- Many factors need to be taken into account
  - Identifying observations that represent the forecast event
    - ✓Example: Precipitation accumulation over an hour at a point
  - For a gridded forecast there are many options for the matching process
    - Point-to-grid
      - Match obs to closest gridpoint
    - Grid-to-point
      - Interpolate?
      - Take largest value?

Matching forecasts and observations

- Point-to-Grid and Grid-to-Point

- Matching approach can impact the results of the verification





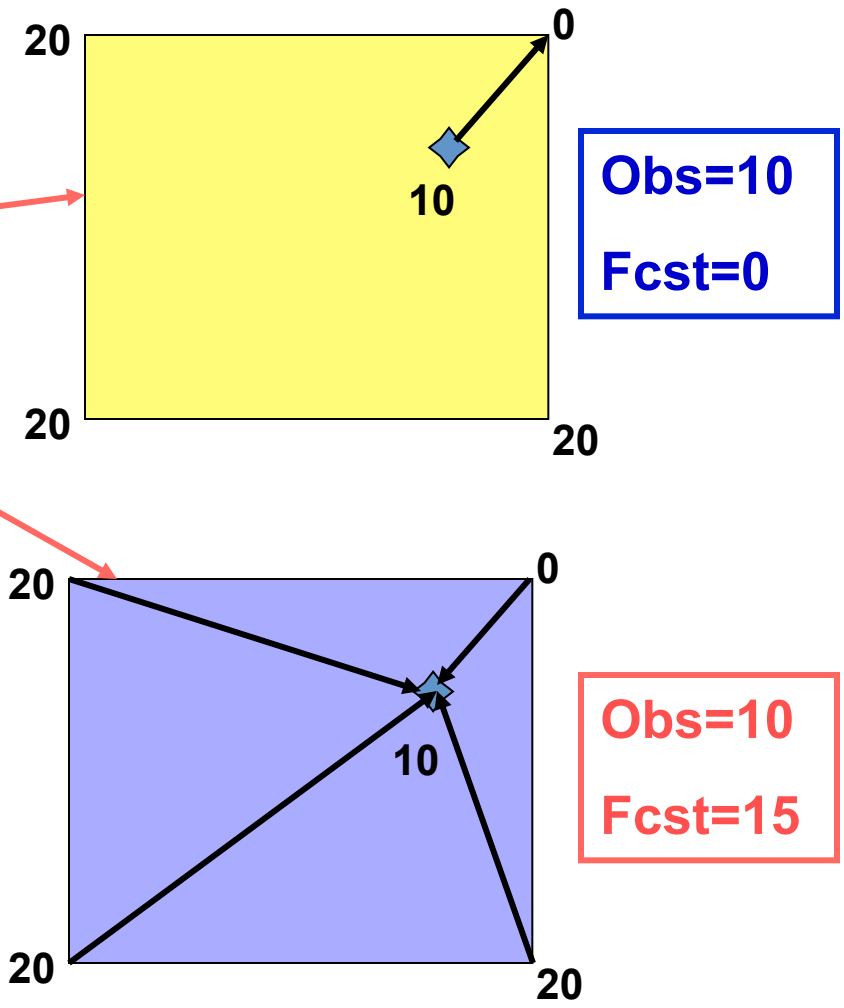
Matching forecasts and observations

Example:

- Two approaches:
  - Match rain gauge to nearest gridpoint ***or***
  - Interpolate grid values to rain gauge location
    - Crude assumption: equal weight to each gridpoint
- Differences in results associated with matching:

"Representativeness" difference

Will impact most verification scores



Matching forecasts and observations

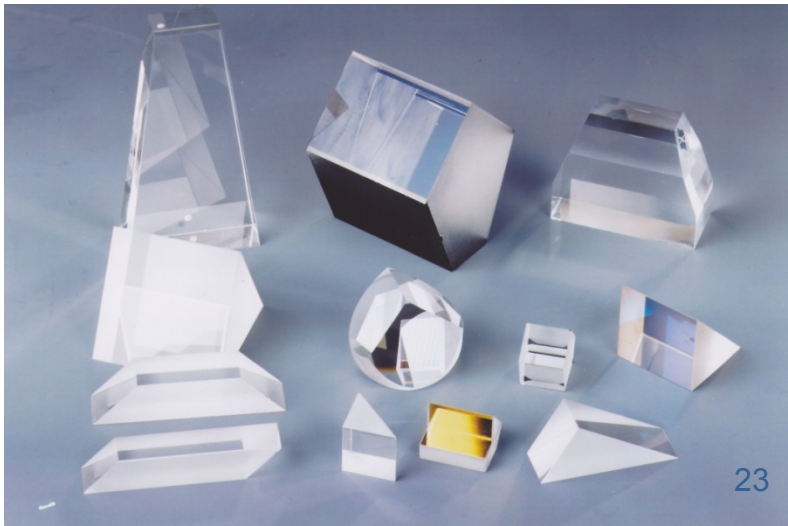
Final point:

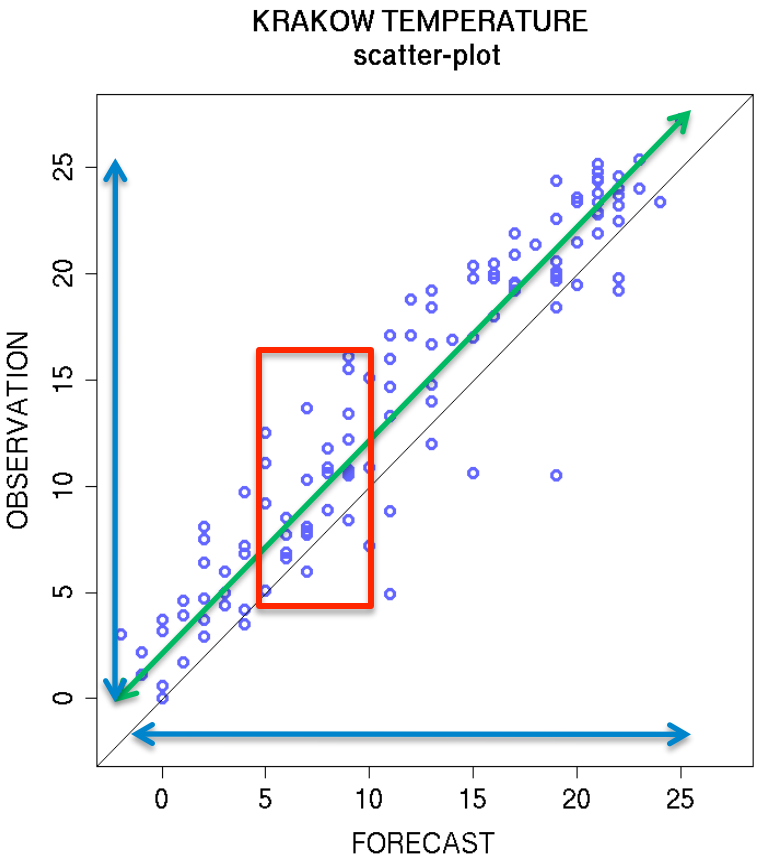
- It is not advisable to use the model analysis as the verification "observation".

- Why not??
- Issue: Non-independence!!

Verification attributes

- Verification attributes measure different aspects of forecast quality
  - Represent a range of characteristics that should be considered
  - Many can be related to joint, conditional, and marginal distributions of forecasts and observations





Joint : The probability of two events in conjunction.

Pr (Tornado forecast AND Tornado observed) = 30 / 2800 = 0.01

Conditional : The probability of one variable given that the second is already determined.

Pr (Tornado Observed | Tornado Fcst) = 30/50 = 0.60

Marginal : The probability of one variable without regard to the other.

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

Pr(Yes Forecast) = 100/2800 = 0.04
Pr(Yes Obs) = 50 / 2800 = 0.02

Verification attribute examples

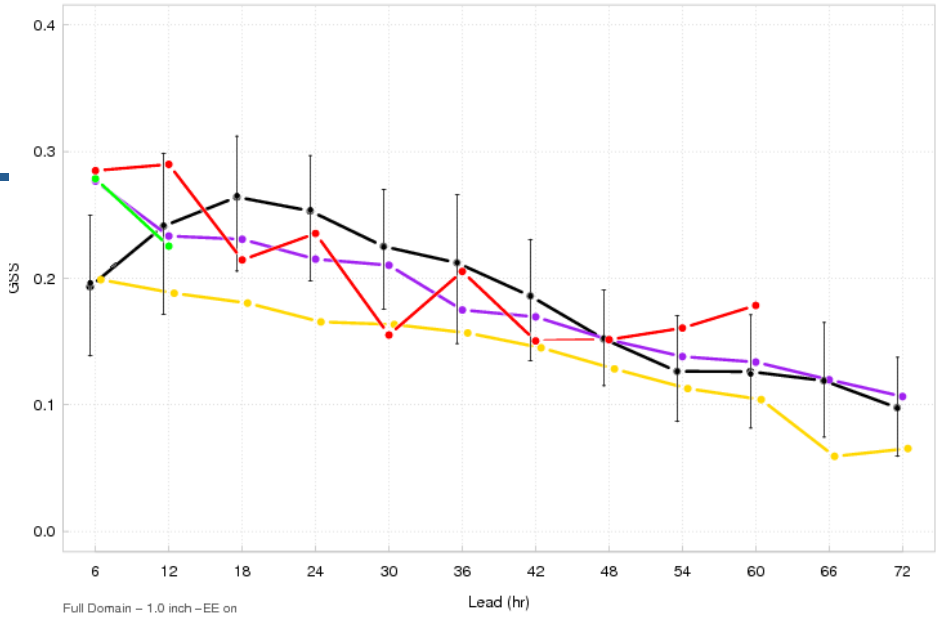
- Bias
  - (Marginal distributions)
- Correlation
  - Overall association (Joint distribution)
- Accuracy
  - Differences (Joint distribution)
- Calibration
  - Measures conditional bias (Conditional distributions)
- Discrimination
  - Degree to which forecasts discriminate between different observations (Conditional distribution)

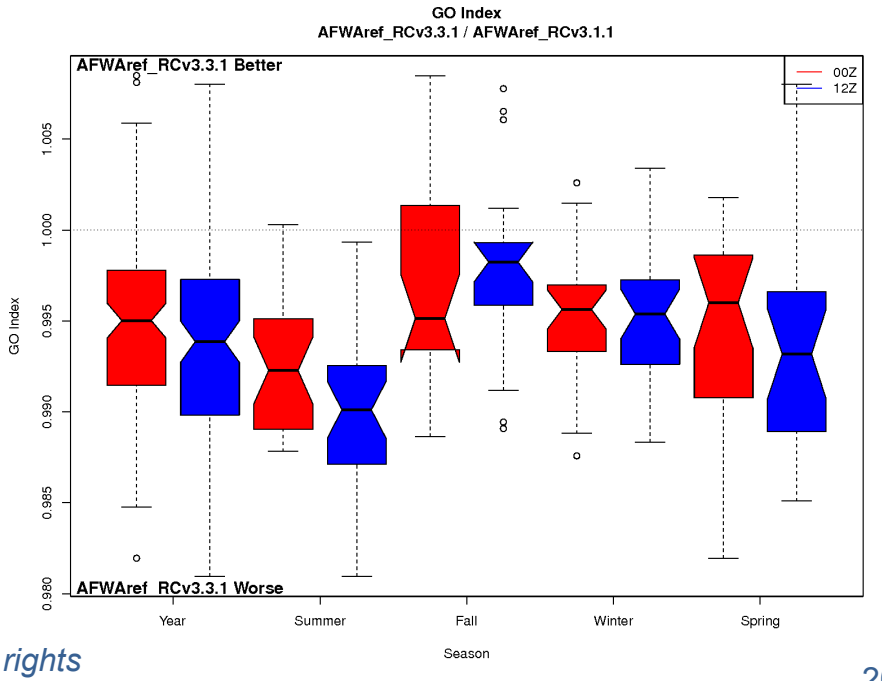
Comparison and inference

- Uncertainty in scores and measures should be estimated whenever possible

- Uncertainty arises from
  - Sampling variability
  - Observation error
  - Representativeness differences

- Erroneous conclusions can be drawn regarding improvements in forecasting systems and models





Miscellaneous issues

- In order to be *verified*, forecasts must be formulated so that they are *verifiable*!
  - Corollary: All forecasts should be verified – if something is worth forecasting, it is worth verifying
- Stratification and aggregation
  - Aggregation can help increase sample sizes and statistical robustness but can also hide important aspects of performance
    - ✓Most common regime may dominate results, mask variations in performance.
  - Thus it is very important to *stratify results into meaningful, homogeneous sub-groups*

Some key things to think about …

Who…
- …wants to know?

What…
- … does the user care about?
- … kind of parameter are we evaluating? What are its characteristics (e.g., continuous, probabilistic)?
- … thresholds are important (if any)?
- … forecast resolution is relevant (e.g., site-specific, area-average)?
- … are the characteristics of the obs (e.g., quality, uncertainty)?
- … are appropriate methods?

Why…
- …do we need to verify it?

Some key things to think about…

How…

- …do you need/want to present results (e.g., stratification/aggregation)?

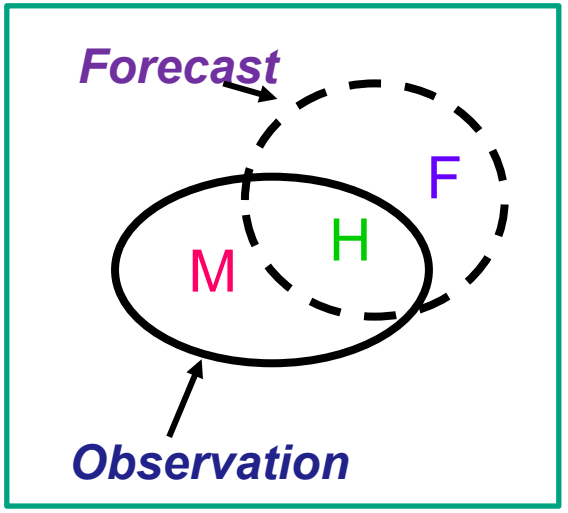
Which…

- …methods and metrics are appropriate?
- … methods are required (e.g., bias, event frequency, sample size)

GSS

CSI

Freq Bias



Categorical Verification

Tara Jensen

Contributions from Matt Pocernich, Eric Gilleland,
Tressa Fowler, Barbara Brown and others

Base Rate

Hit Rate

FAR

POD Y

Finley Tornado Data (1884)

LIEUTENANT JOHN P. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.

Forecast answering the question:

Will there be a tornado?

YES

NO

Observation answering the question:

Did a tornado occur?

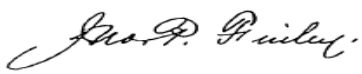
YES

NO

Answers fall into 1 of 2 categories ⁂ Forecasts and Obs are Binary

Finley Tornado Data (1884)

LIEUTENANT JOHN P. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.



		Observed		
Forecast		Yes	No	Total
	Yes			
	No			
	Total			

Contingency Table

A Success?

LIEUTENANT JOHN P. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.

		Observed		
Forecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

Percent Correct = (28+2680)/2803 = 96.6% !!!!

What if forecaster never forecasted a tornado?

LIEUTENANT JOHN P. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.

John P. Finley.

		Observed		
Forecast		Yes	No	Total
	Yes	0	0	0
	No	51	2752	2803
	Total	51	2752	2803

Percent Correct = (0+2752)/2803 = 98.2% !!!!

maybe Accuracy is not the most informative statistic

But the contingency table concept is good…

2 x 2 Contingency Table

		Observed		
Forecast		Yes	No	Total
	Yes	Hit	False Alarm	Forecast Yes
	No	Miss	Correct Negative	Forecast No
	Total	Obs. Yes	Obs. No	Total

Example: Accuracy = (**Hits**+**Correct Negs**)/**Total**

MET supports both 2x2 and NxN Contingency Tables

Common Notation

(however not universal notation)

		Observed		
Forecast		Yes	No	Total
	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	n

Example: Accuracy = (a+d)/n

What if data are not binary?

Threshold

Temperature < 0 C
Precipitation > 1 inch
CAPE > 1000 J/kg
Ozone > 20 μg/m³
Winds at 80 m > 24 m/s
500 mb HGTS < 5520 m
Radar Reflectivity > 40 dBZ
MSLP < 990 hPa
LCL < 1000 ft
Cloud Droplet Concentration > 500/cc

Hint: Pick a threshold that is meaningful to your end-user

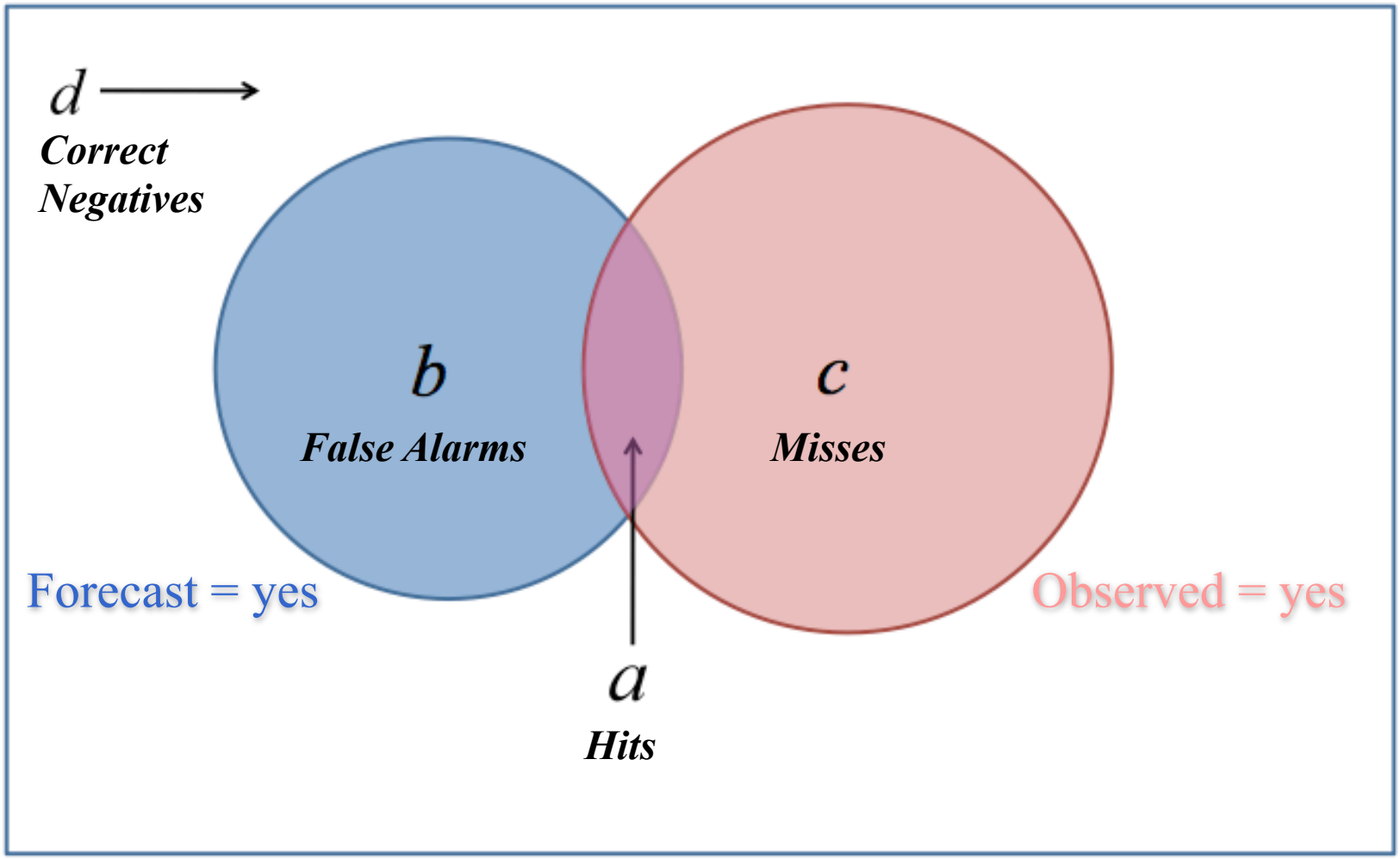
Contingency Table for Freezing Temps (i.e. T<=0 C)

		Observed		
Forecast		<= 0C	> 0C	Total
	<= 0C	a	b	a+b
	> 0C	c	d	c+d
	Total	a+c	b+d	n

Another Example:

Base Rate (aka sample climatology) = (a+c)/n

Alternative Perspective on Contingency Table



Conditioning to form a statistic

- Considers the probability of one event given another event
- Notation: p(X|Y=1) is probability of X occuring given Y=1 or in other words Y=yes

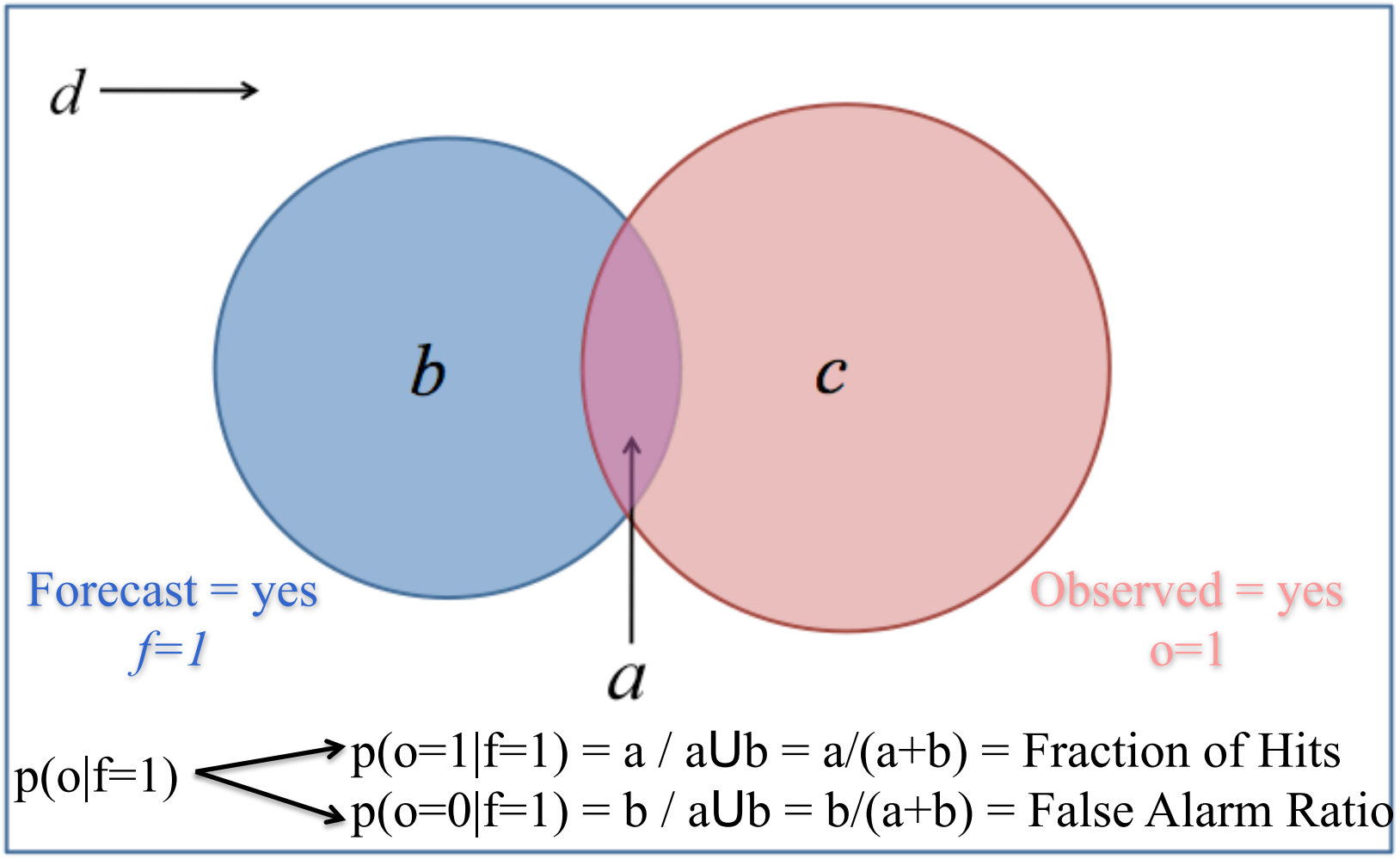
Conditioning on Fcst provides:

- Info about how your forecast is performing
- Apples-to-Oranges comparison if comparing stats from 2 models

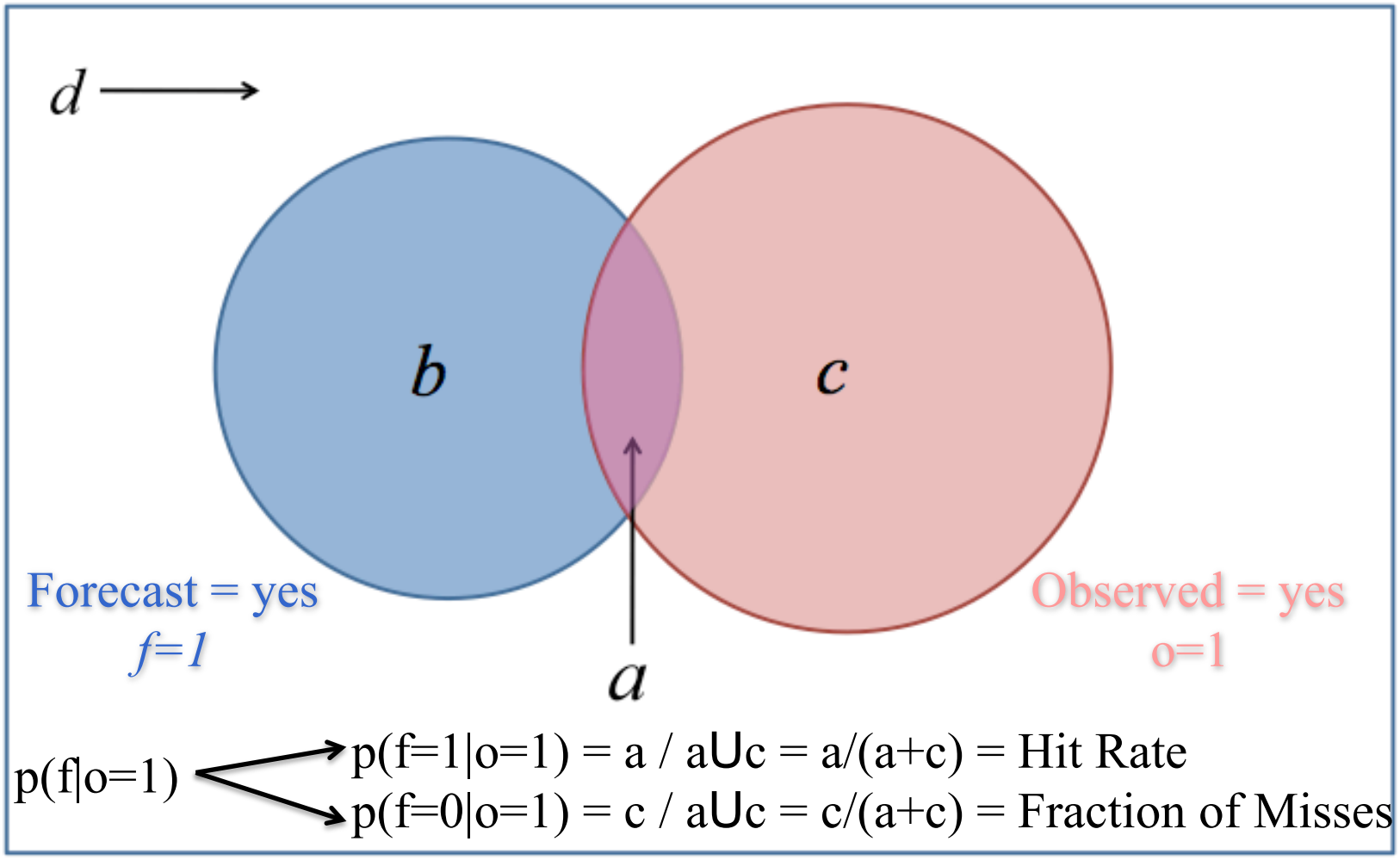
Conditioning on Obs provides:

- Info about ability of forecast to discriminate between event and non-event - also called Conditional Probability or "Likelihood"
- Apples-to-Apples comparison if comparing stats from 2 models

Conditioning on forecasts



Conditioning on observations



What's considered good?

Conditioning on Forecast

Fraction of hits - $p(f=1|o=1) = a/(a+b)$: close to 1

False Alarm Ratio - $p(f=0|o=1) = b/(a+b)$: close to 0

Conditioning on Observations

Hit Rate - $p(f=1|o=1) = a/(a+c)$: close to 1

[aka Probability of Detection Yes (PODy)]

Fraction of misses $p(f=0|o=1) = a/(a+c)$: close to 0

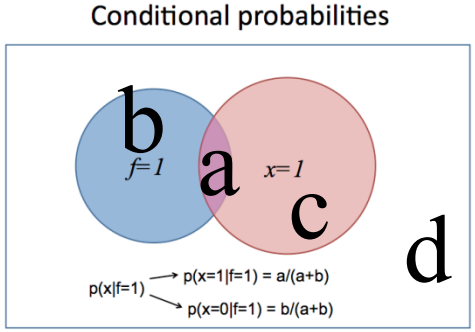
Examples of Categorical Scores

(most based on conditioning)

- Hit Rate (PODy) = a/(a+c)
- PODn = d/(b+d) = (1 – POFD)
- False Alarm Rate (POFD) = b/(b+d)
- False Alarm **Ratio** (FAR) = b/(a+b)
- (Frequency) Bias (FBIAS) = (a+b)/(a+c)
- Threat Score or Critical Success Index = a/(a+b+c) (CSI)

POD
Probability of Detection

POFD
Probability of False Detection



		Observed		
Forecast		Yes	No	Total
	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	n

Examples of Contingency table calculations

		Observed		
Forecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

Threat Score = 28 / (28 + 72+ 23) = 0.228

Probability of Detection = 28 / (28 + 23) = 0.55

False Alarm Ratio= 72/(28 + 72) = 0.720

Skill Scores

How do you compare the skill of easy to predict events with difficult to predict events?

- Provides a single value to summarize performance.
- Reference forecast - best naive guess; persistence; climatology.
- Reference forecast must be comparable.
- Perfect forecast implies that the object can be perfectly observed.

Generic Skill Score

$$SS = \frac{(A - A_{ref})}{(A_{perf} - A_{ref})}$$

where A = any measure
ref = reference
perf = perfect

Example: $$MSESS = 1 - \frac{MSE}{MSE_{climo}}$$

where MSE = Mean Square Error

Interpreted as fractional improvement over reference forecast

- Reference could be: Climotology, Persistence, your baseline forecast, etc..
- Climotology could be a separate forecast or a gridded forecast sample climatology

SS typically positively oriented with 1 as optimal

Commonly Used Skill Scores

- **Gilbert Skill Score** - based on the CSI corrected for the *number of hits that would be expected by chance*.
- **Heidke Skill Score** - based on Accuracy corrected by the *number of hits that would be expected by chance*.
- **Hanssen-Kuipers Discriminant** – (Pierce Skill Score) measures the ability of the forecast to discriminate between (or correctly classify) events and non-events. H-K=POD-POFD

- **Brier Skill Score** for probabilistic forecasts
- **Fractional Skill Score** for neighborhood methods
- **Intensity-Scale Skill Score** for wavelet methods

Empirical ROC

ROC – Receiver Operating Characteristic

Used to determine how well forecast discriminates between event and non-event.

How to construct:

- Bin your data
- Calculate PODY and POFD by moving thru bins and thus changing the definition of a –d
- Plot using scatter plot

Typically used for Probability Forecasts but can be used any data that has been put into bins

Technique allows non-calibrated (no bias correction) to be compared because it inherently removes model bias from comparison

Example Tables

Binned Continuous Forecast

Fcst 80m Winds (m/s)	# Yes Obs	# No Obs
0-3	146	14
4-6	16	8
7-9	12	3
10-12	10	10
13-15	15	5
16-18	4	9
19-21	7	9
22-24	2	8
25-28	7	8
29<	6	32

Probability Winds will be below Cut-Out Speed

Binned Probabilistic Forecast

PROB	# YES	# NO
0.05	6	32
0.15	7	8
0.25	2	8
0.35	7	9
0.45	4	9
0.55	15	5
0.65	10	10
0.75	12	3
0.85	16	8
0.95	146	14

Mid-points

Calculation of Empirical ROC

Used to determine how well forecast discriminates between event and non-event.

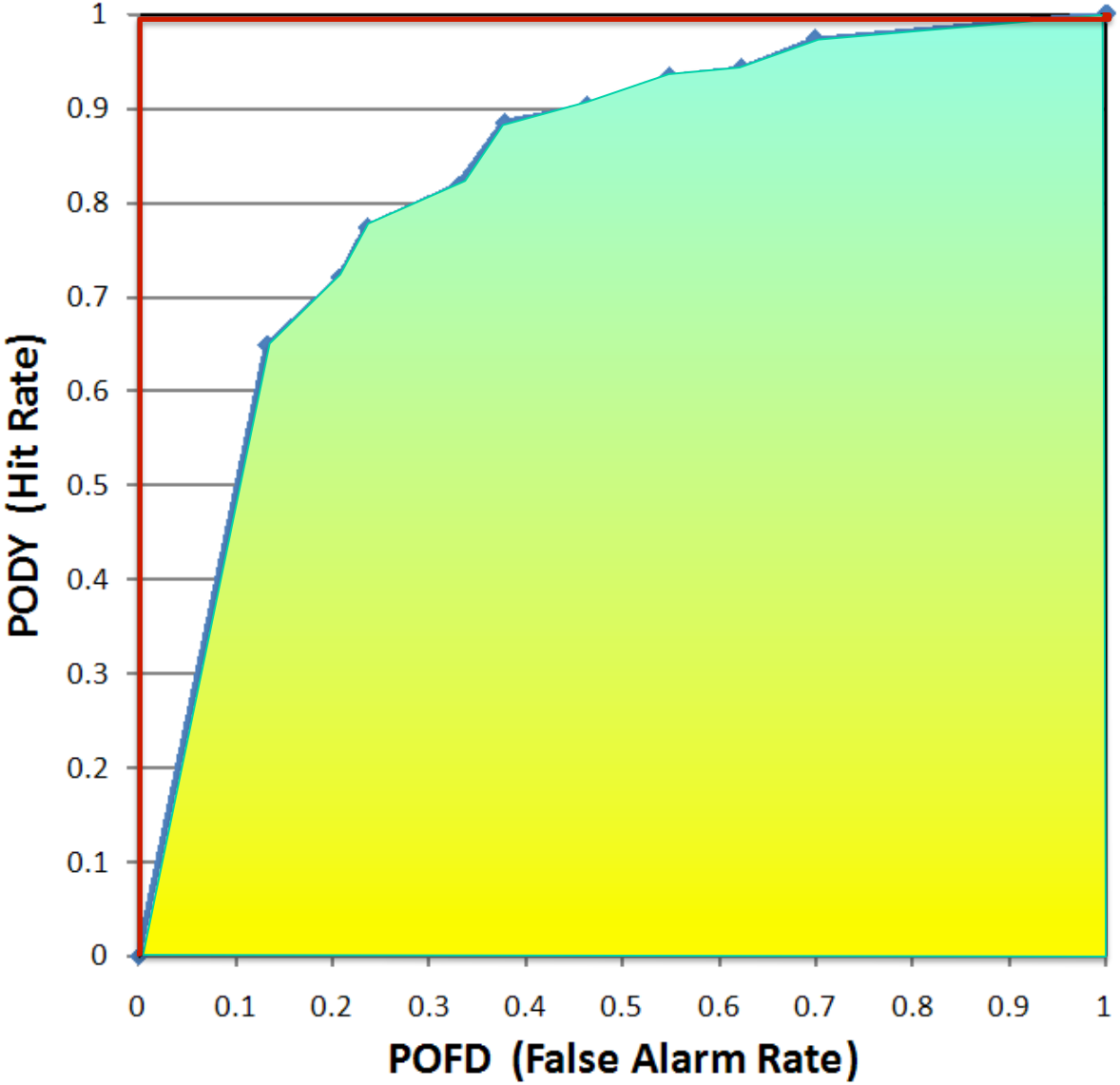
PODY	POFD
Hit Rate	vs. False Alarm Rate
0.98	0.55
0.90	0.46
0.88	0.38

Does not need to be a probability!

Does not need to be calibrated!

PROB	# YES	# NO
0.05	6	32
0.15	c 7	8 d
0.25	15 2	8 48
0.35	22 7	9 57
0.45	26 4	9 63
0.55	15	5
0.65	a 10	10 b
0.75	210 12	3 58
0.85	203 16	8 49
0.95	199 146	14 40

Empirical ROC Curve



Perfect

Diagonal line represents No Skill
(hit just as likely as a false alarm)

If line fall under Diagonal
Fcst Worse than Random Guess

Area under the ROC curve is a useful measure **(AUC)**

Perfect = 1, Random = 0.5

Verification of Continuous Forecasts

Presented by

Barbara G. Brown

Adapted from presentations created by
Barbara Casati and Tressa Fowler

DTC
Developmental Testbed Center

- Exploratory methods
  - Scatter plots
  - Discrimination plots
  - Box plots

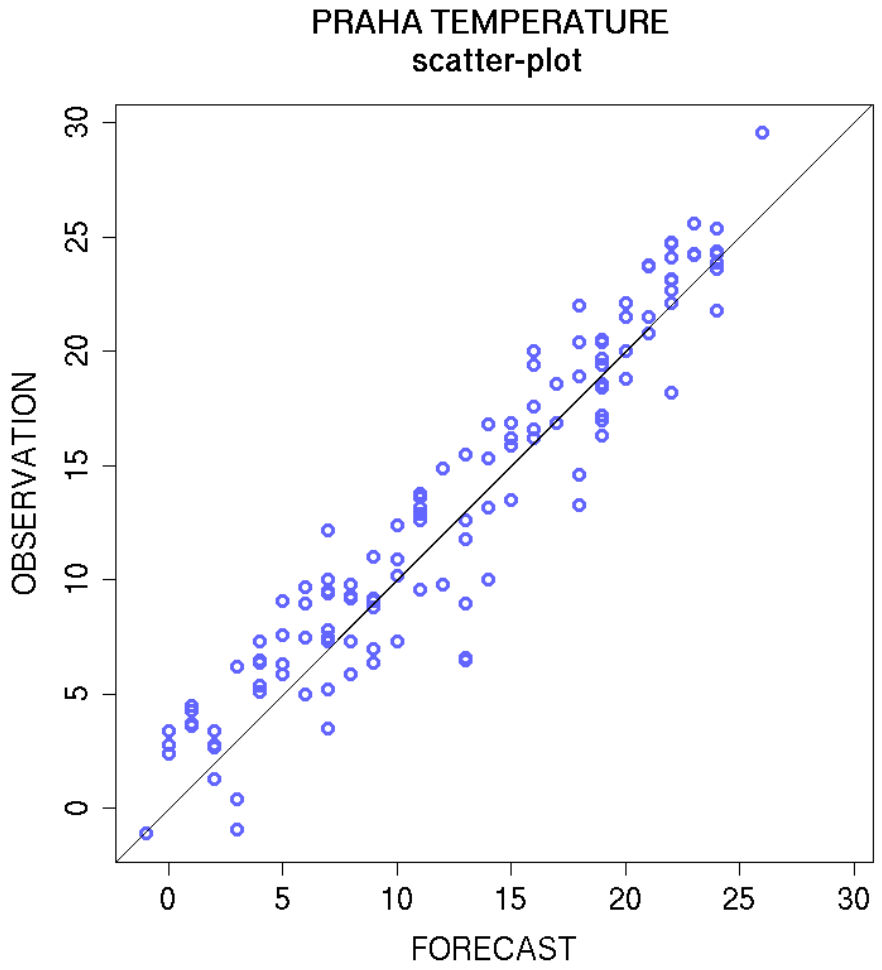
- Statistics
  - Bias
  - Error statistics
  - Robustness
  - Comparisons

Exploratory methods: joint distribution

Scatter-plot: plot of observation versus forecast values

Perfect forecast = obs, points should be on the 45^{o} diagonal

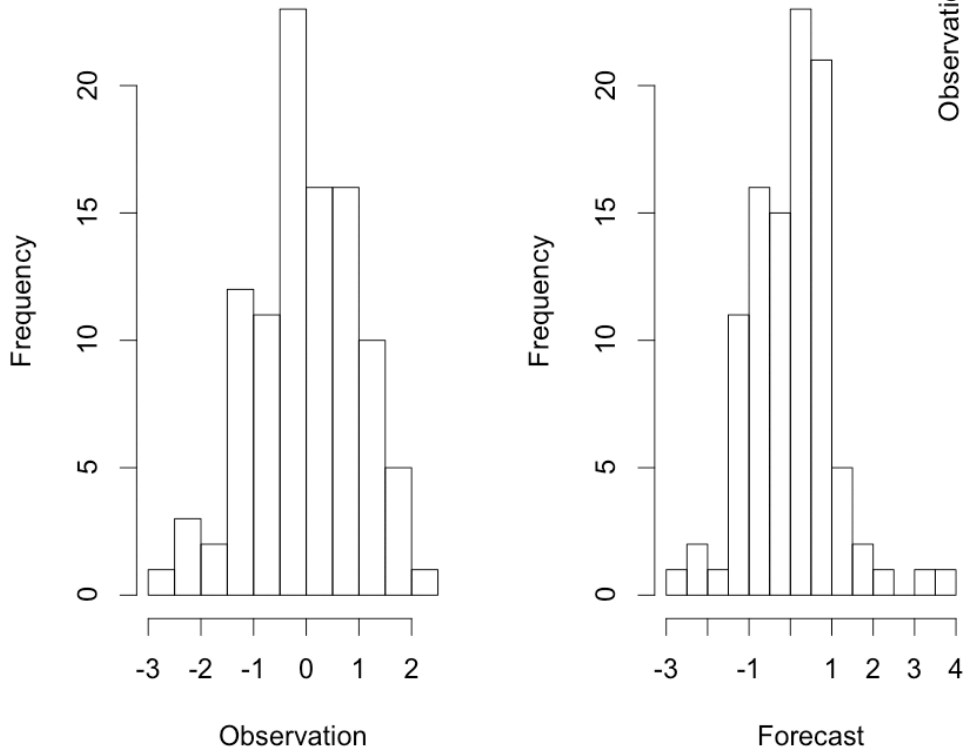
Provides information on: bias, outliers, error magnitude, linear association, peculiar behaviours in extremes, misses and false alarms (link to contingency table)

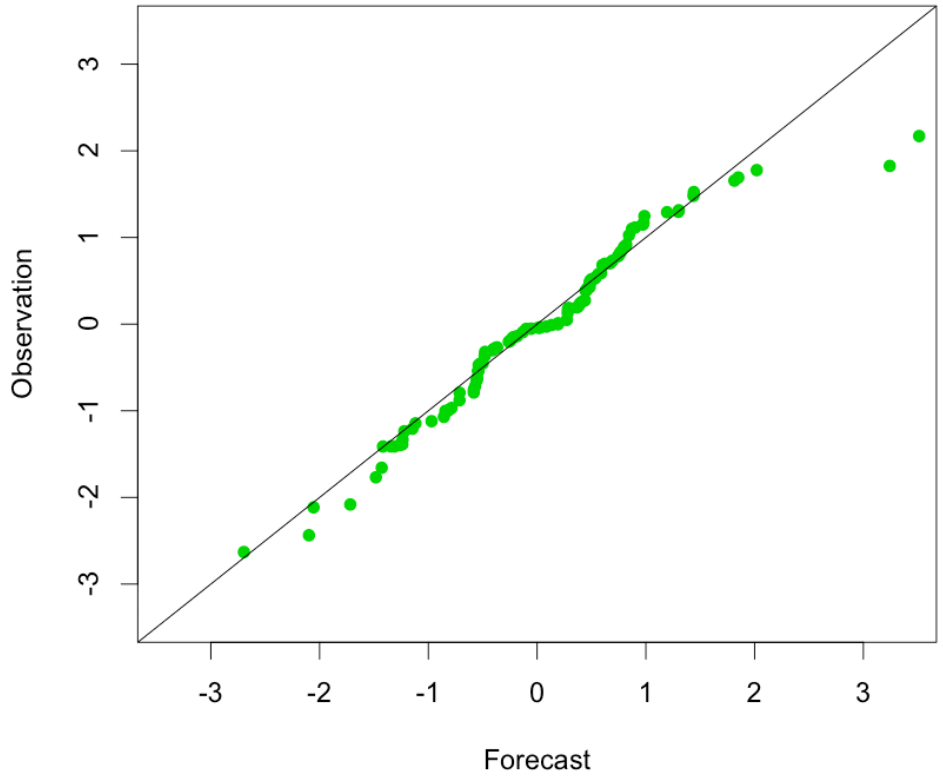


Exploratory methods: marginal distribution

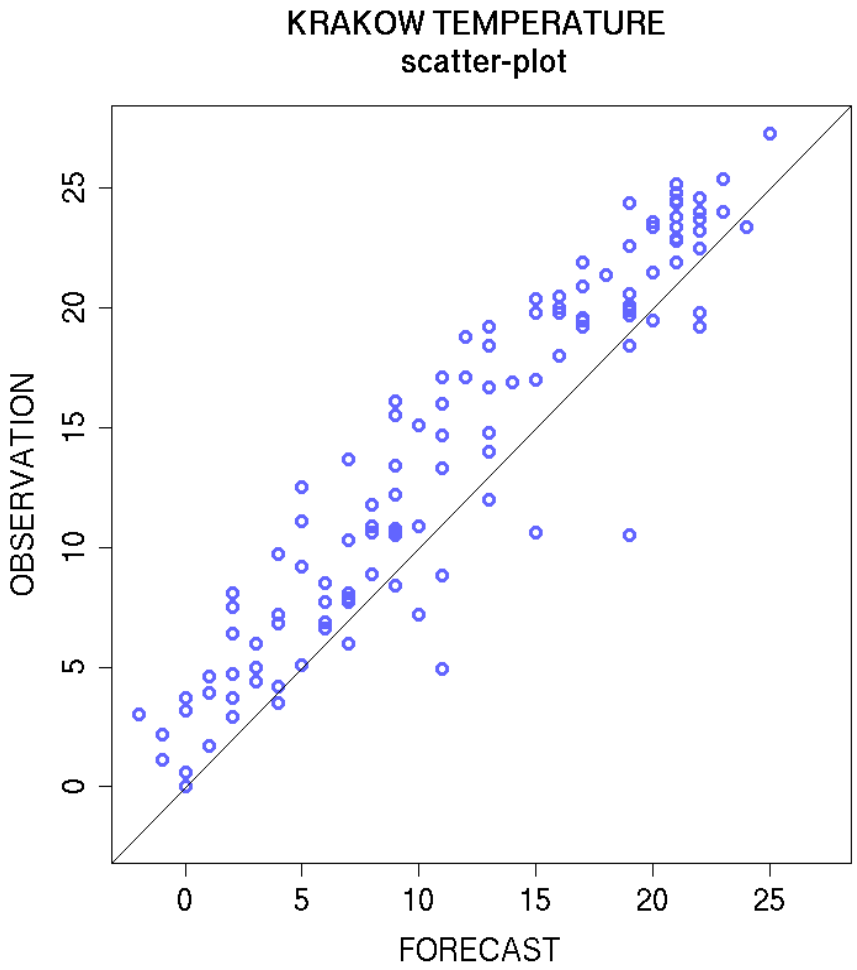
Quantile-quantile plots:

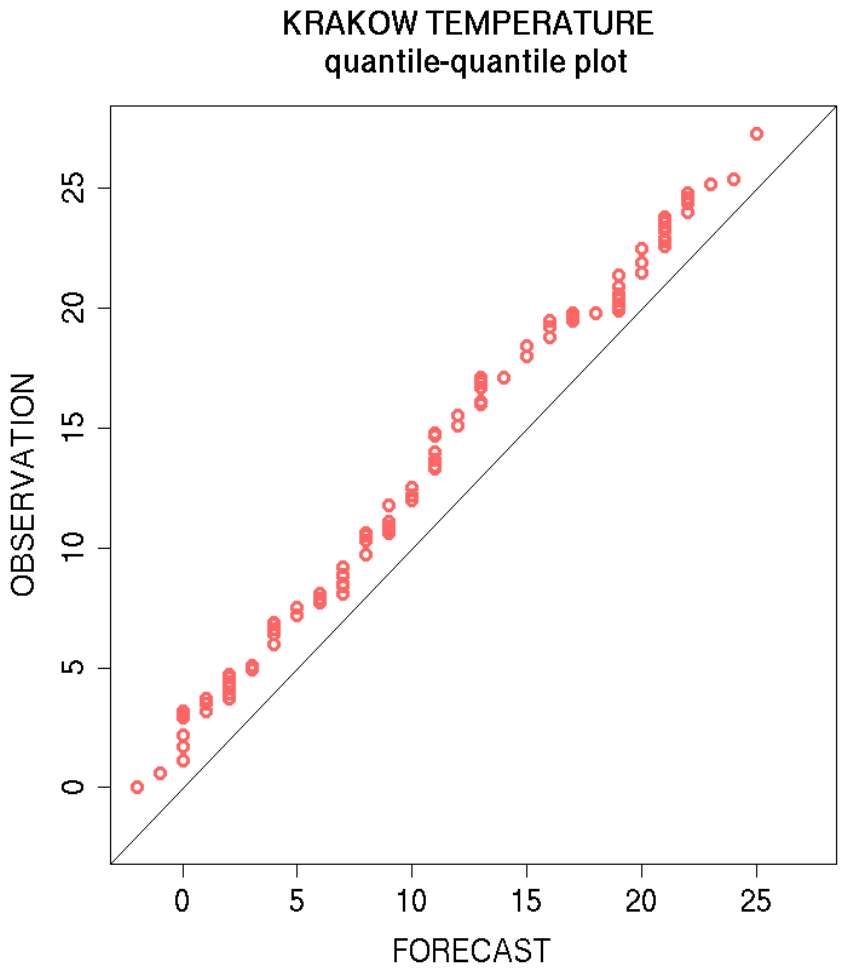
OBS quantile versus the corresponding FRCS quantile





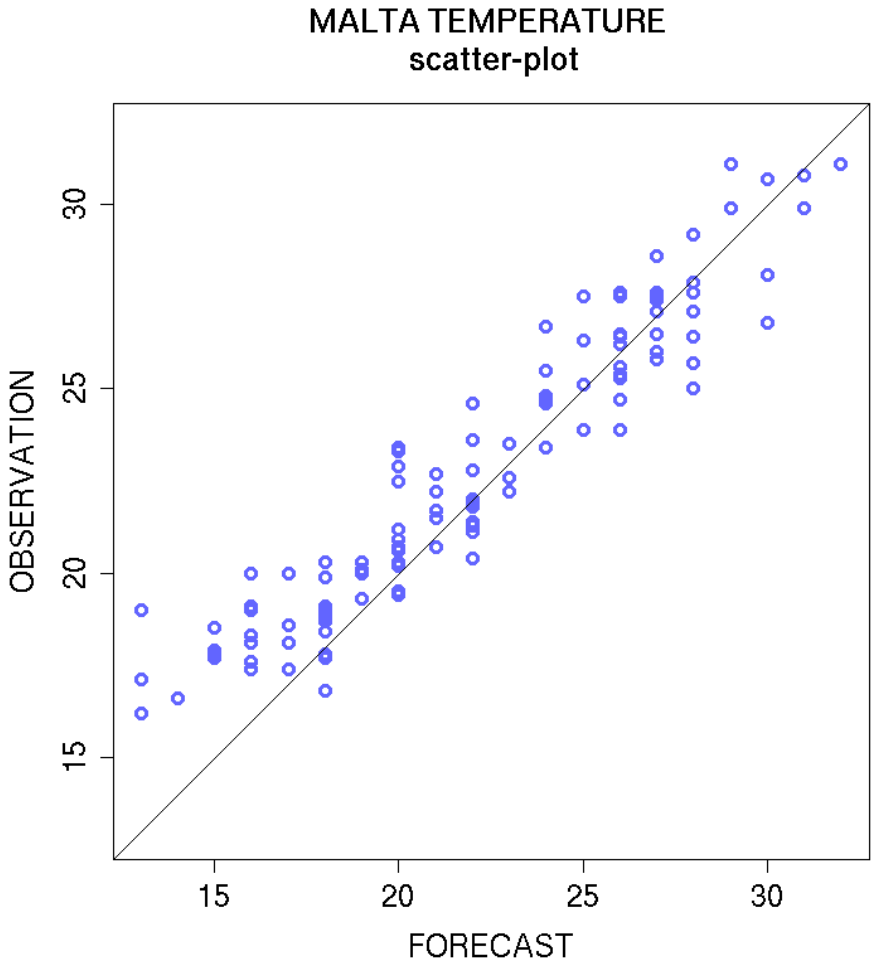
Scatter-plot and qq-plot: example 1
Q: is there any bias? Positive (over-forecast) or negative (under-forecast)?

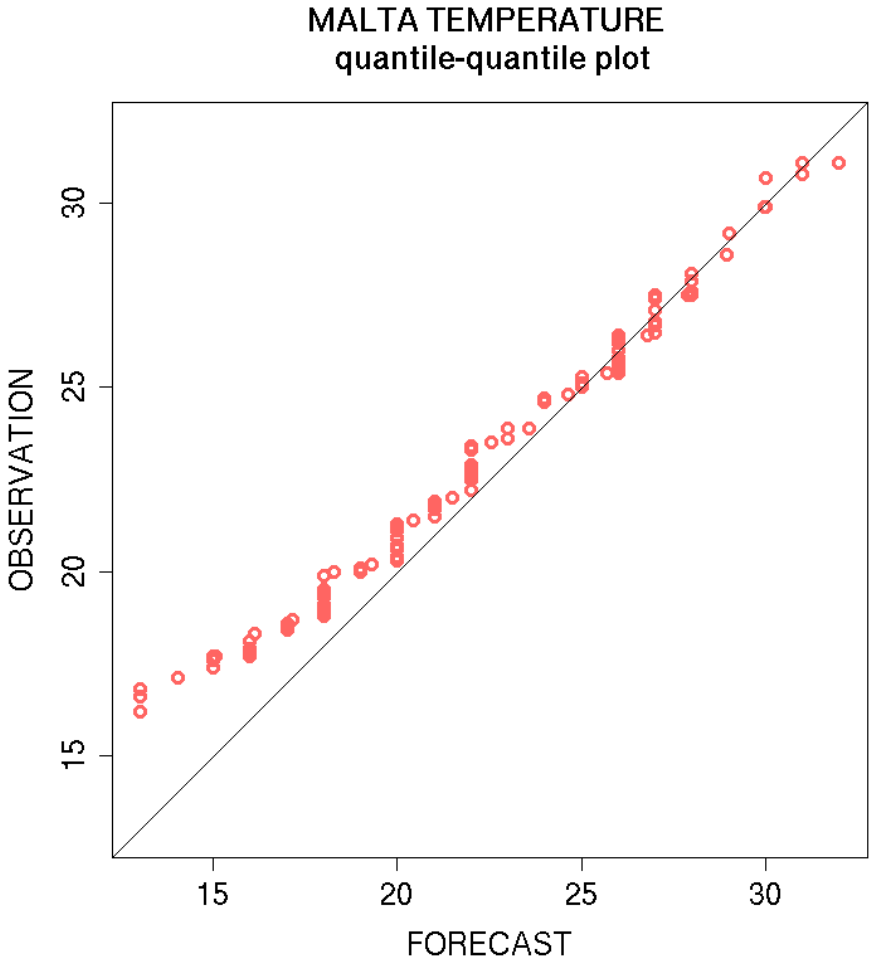




Scatter-plot and qq-plot: example 2

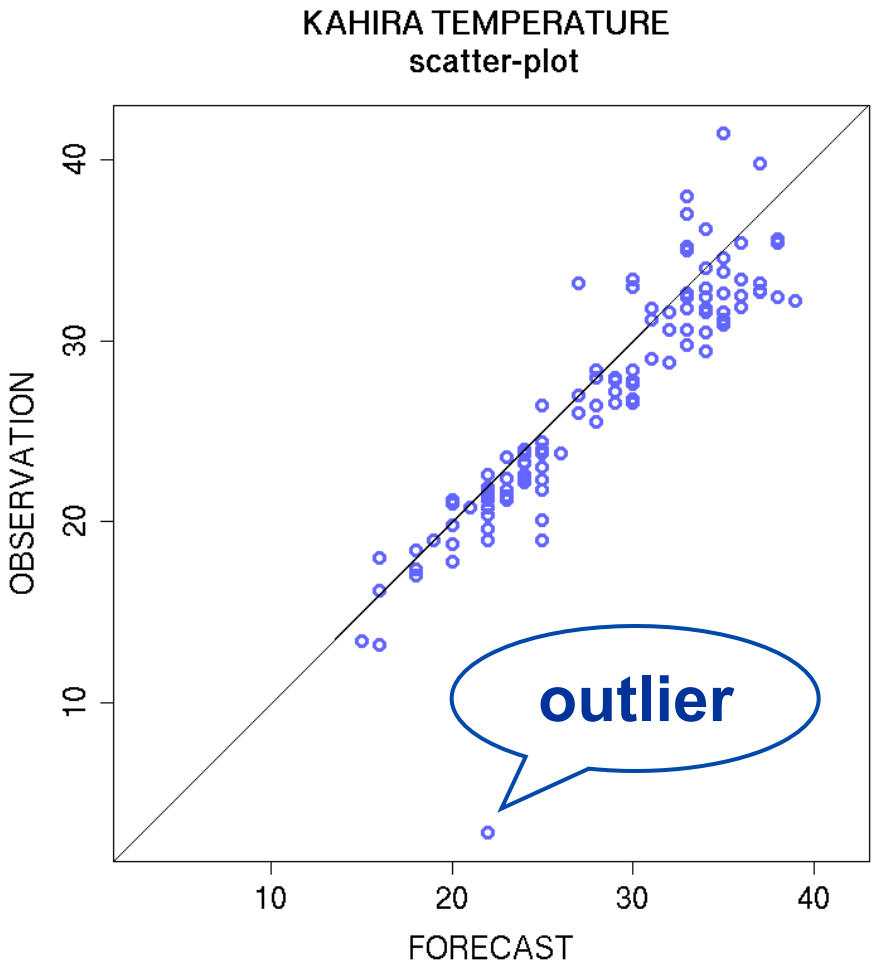
Describe the peculiar behaviour of low temperatures

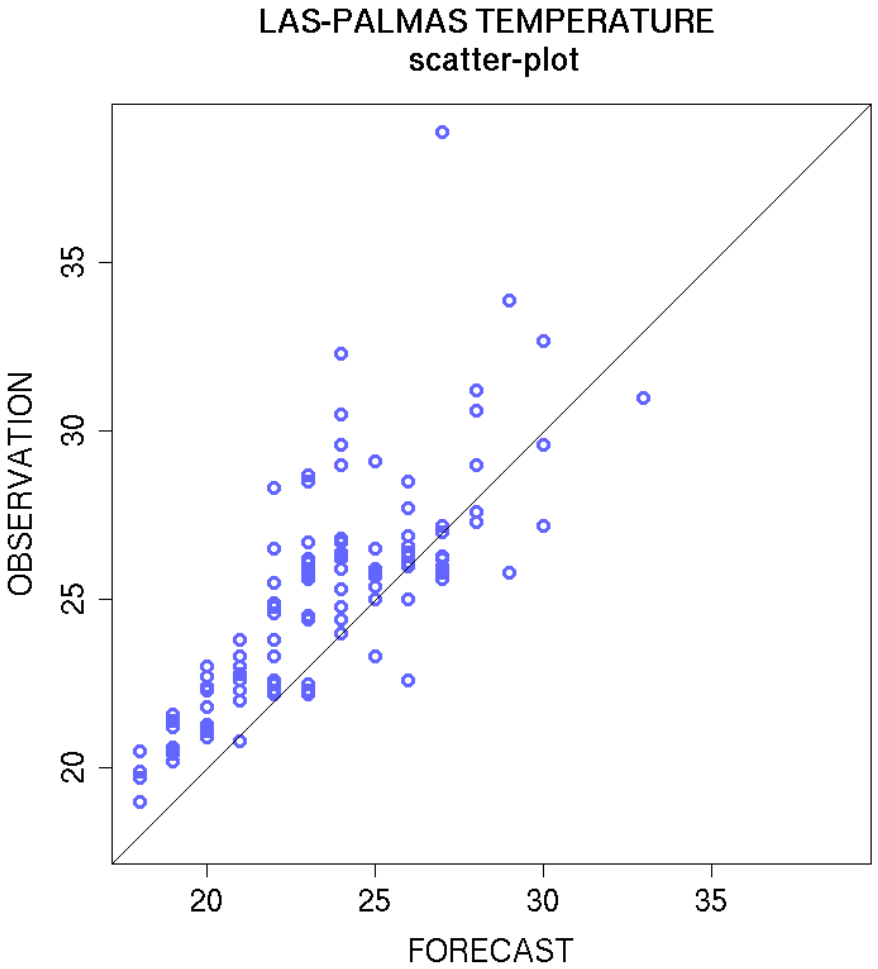




Scatter-plot: example 3

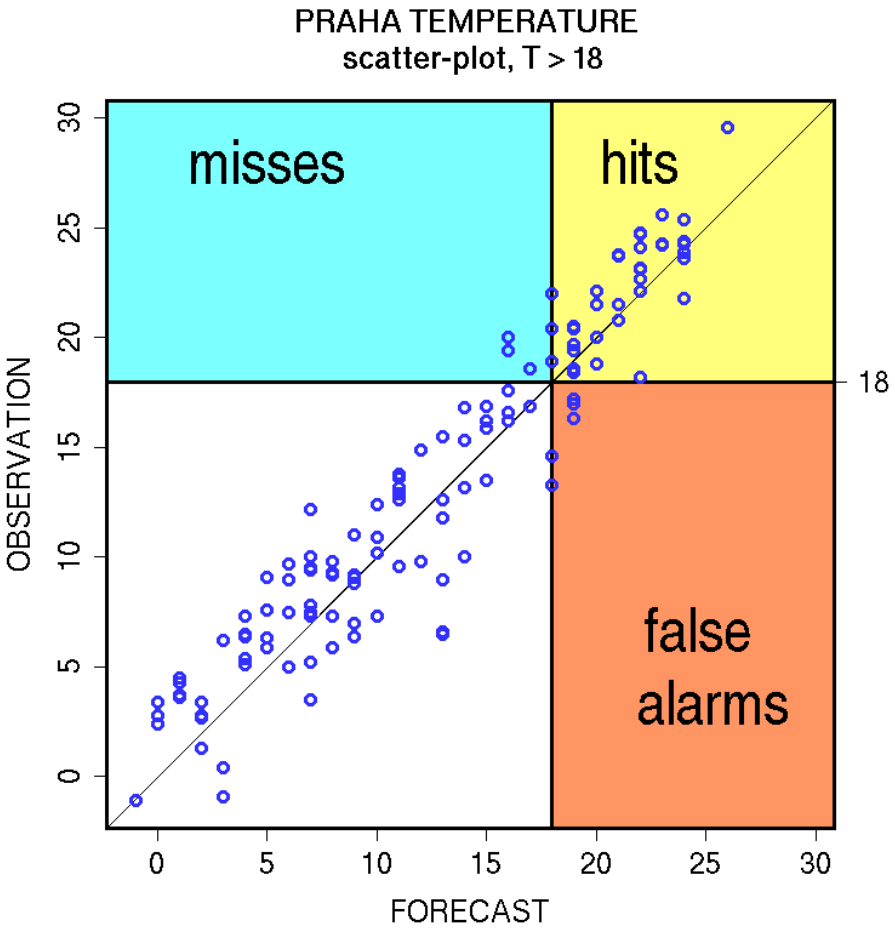
Describe how the error varies as the temperatures grow



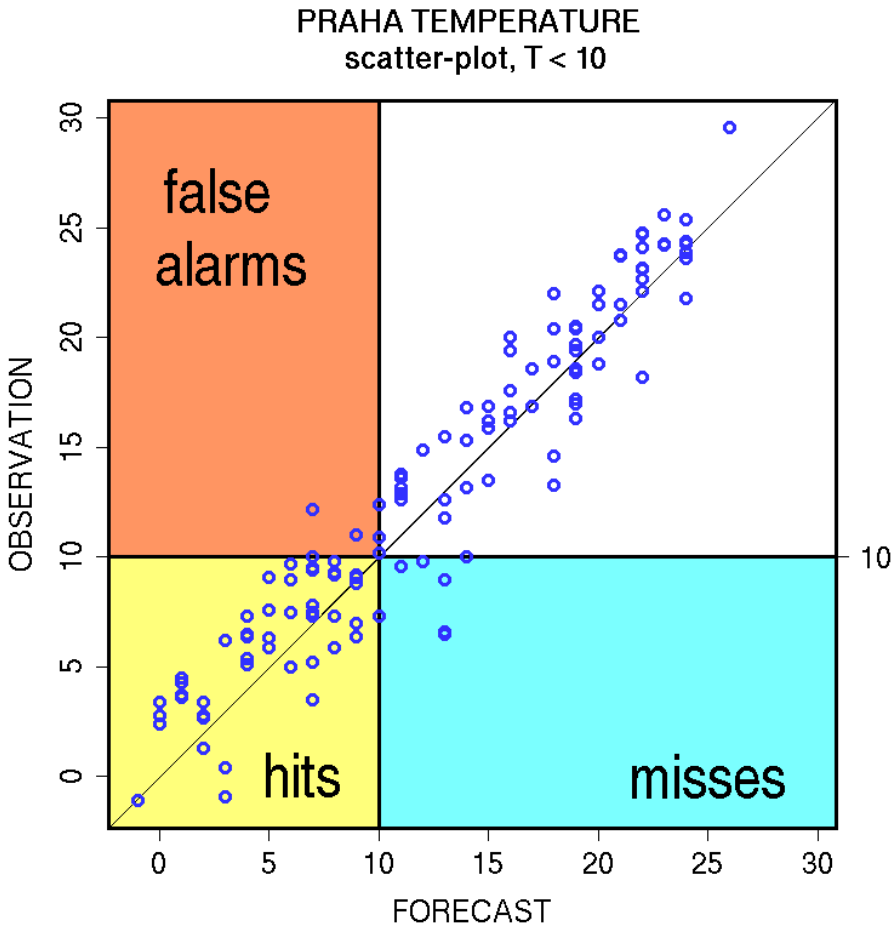


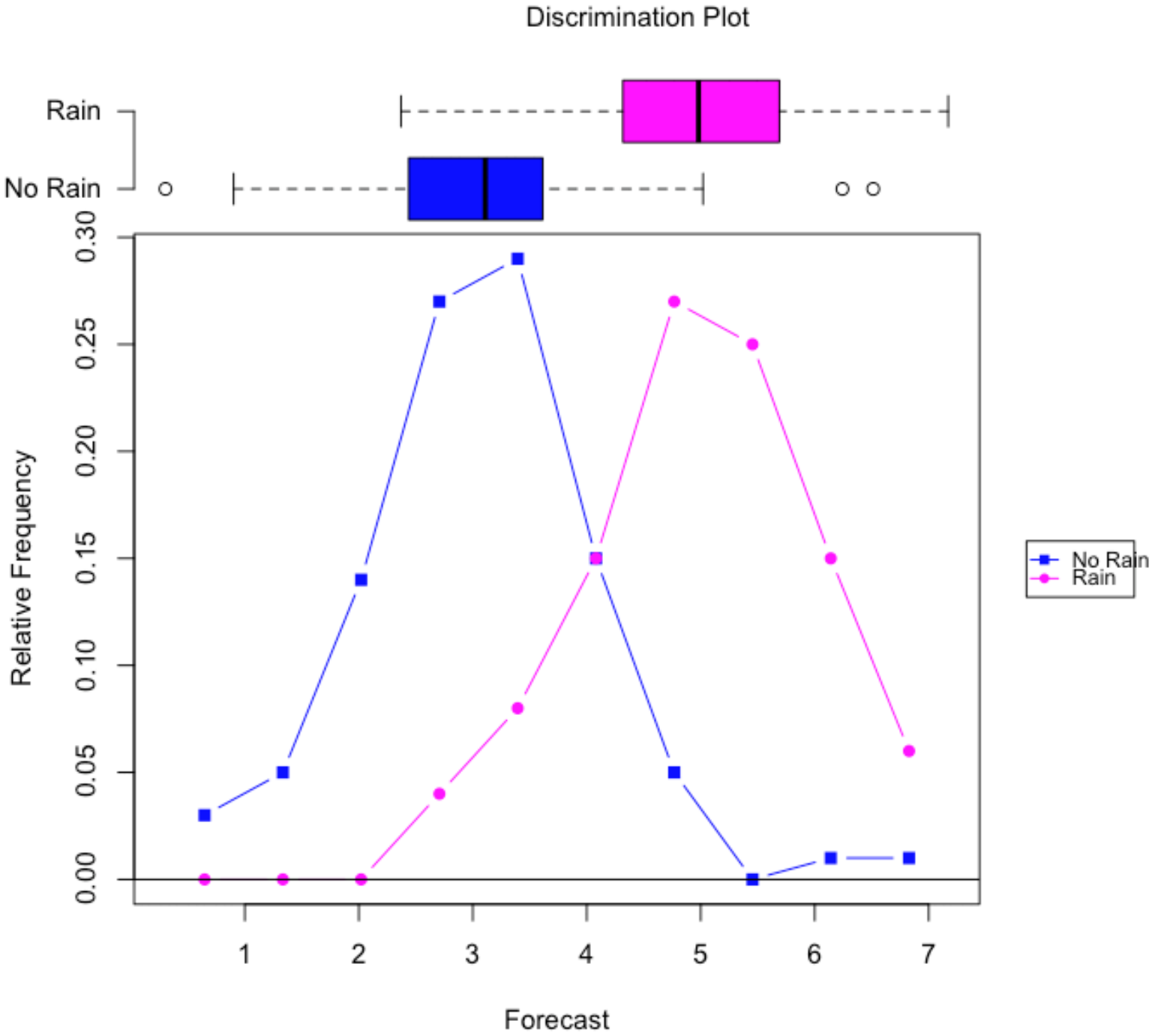
Scatter-plot and Contingency Table

Does the forecast detect correctly temperatures above 18 degrees ?

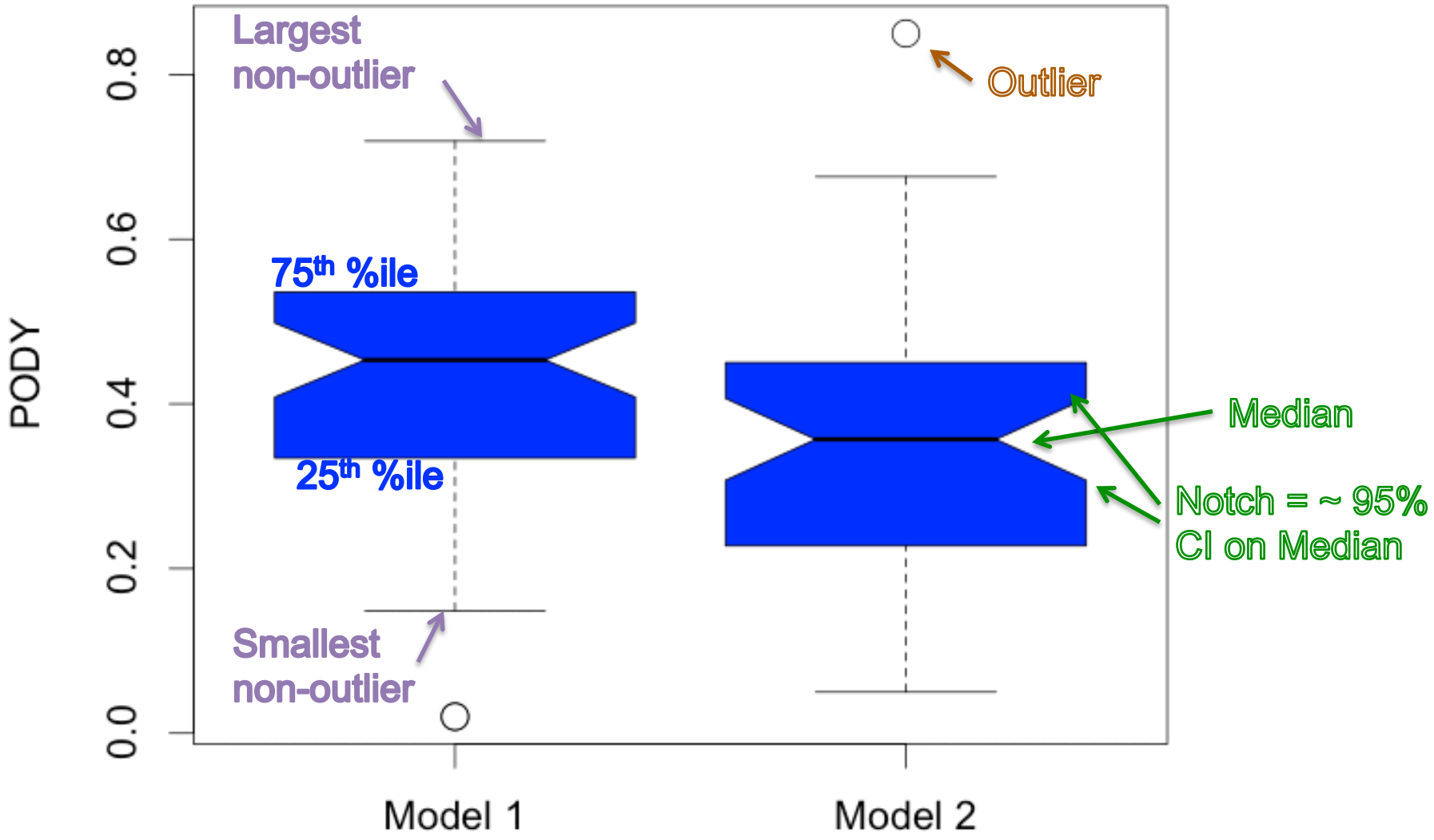


Does the forecast detect correctly temperatures below 10 degrees ?




Discrimination Plot
Rain
No Rain
Relative Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
6
7
Forecast
No Rain
Rain

Example Box (and Whisker) Plot



Exploratory methods: marginal distributions

Visual comparison:
Histograms, box-plots, …

Summary statistics:

- ***Location:***

$$\text{mean} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

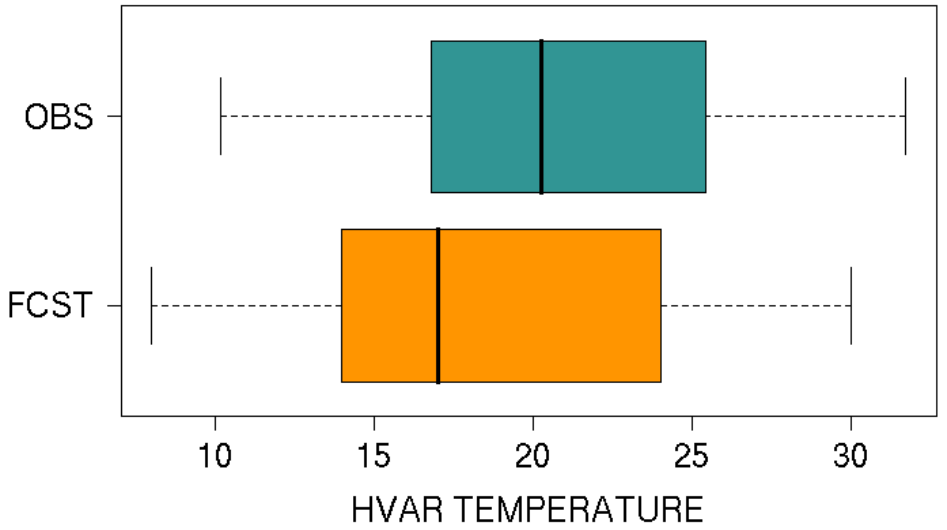
$$\text{median} = q_{0.5}$$

- ***Spread:***

$$\text{st dev} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

$$\text{Inter Quartile Range} =$$

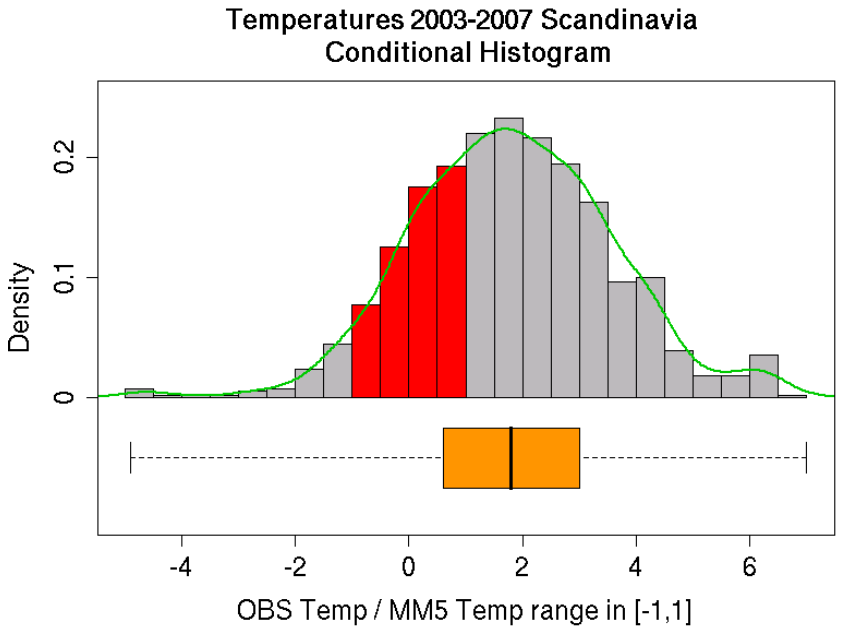
$$\text{IQR} = q_{0.75} - q_{0.25}$$

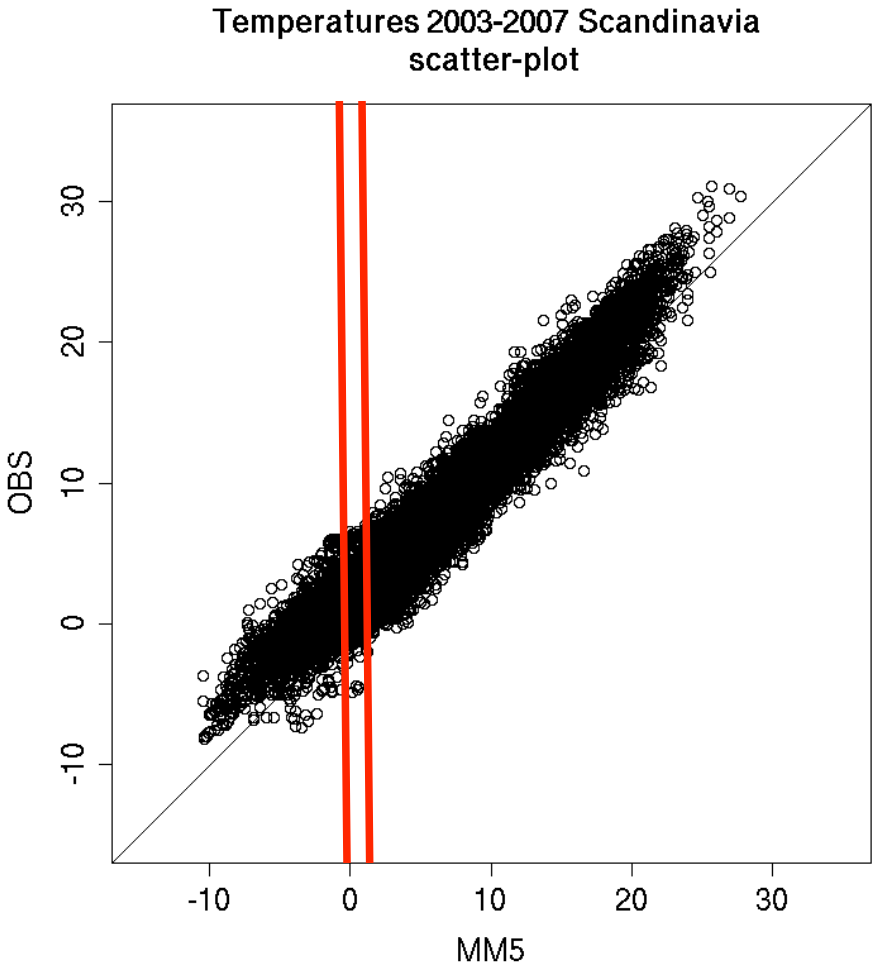


	MEAN	MEDIAN	STDEV	IQR
OBS	20.71	20.25	5.18	8.52
FRCS	18.62	17.00	5.99	9.75

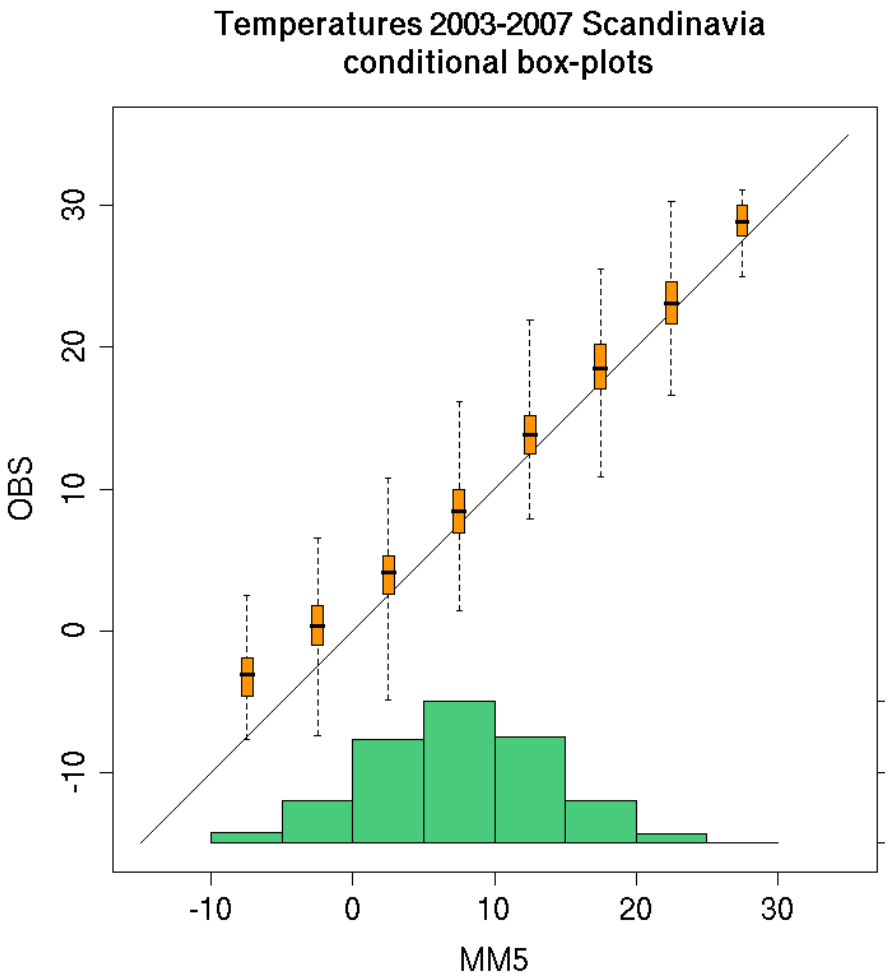
Exploratory methods: conditional distributions

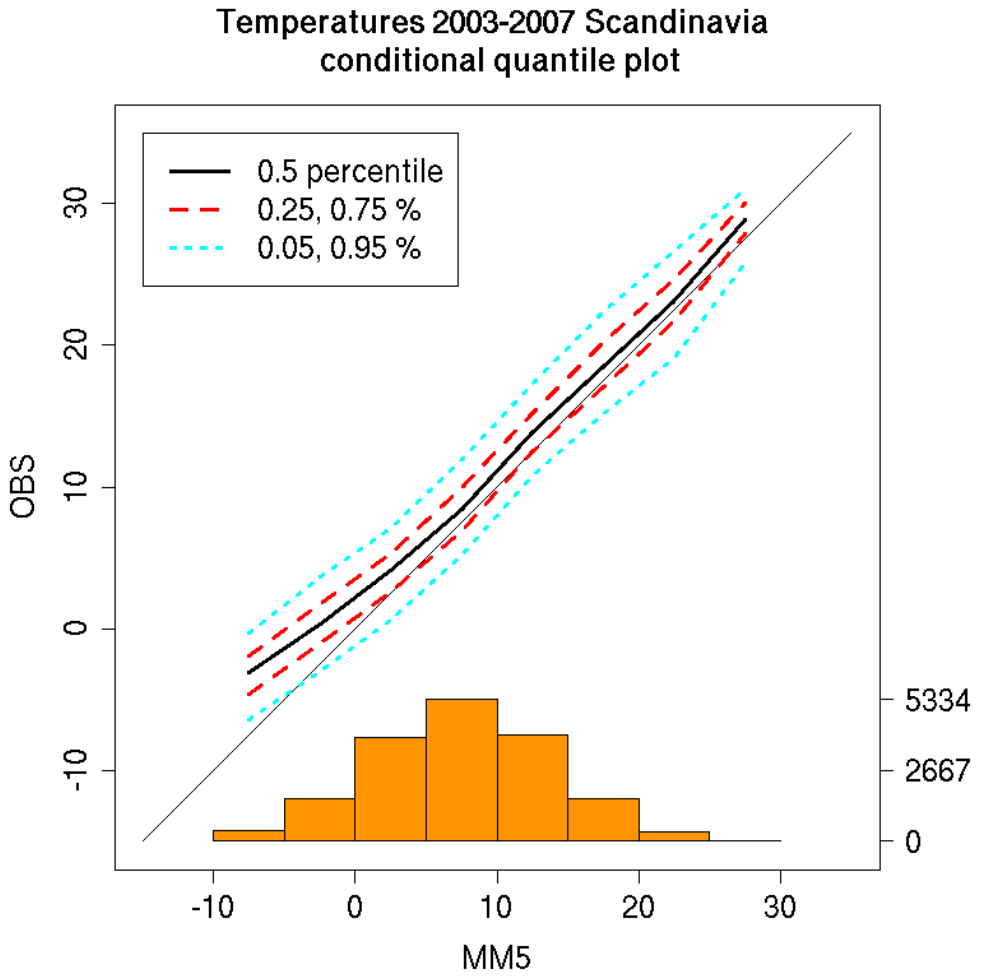
Conditional histogram and conditional box-plot





Exploratory methods: conditional qq-plot





Continuous scores: linear bias

$$\text{linear bias} = \text{Mean Error} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right) = \bar{f} - \bar{o}$$

Attribute: measures the bias

Mean Error = average of the errors = difference between the means

It indicates the average direction of error: positive bias indicates over-forecast, negative bias indicates under-forecast (y=forecast, x=observation)

Does not indicate the magnitude of the error (positive and negative error can cancel outs)

Bias correction: misses (false alarms) improve at the expenses of false alarms (misses). Q: If I correct the bias in an over-forecast, do false alarms grow or decrease ? And the misses ?

Good practice rules: sample used for evaluating bias correction should be consistent with sample corrected (e.g. winter separated by summer); for fair validation, cross validation should be adopted for bias corrected forecasts

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - o_i\right|$$

Attribute: measures accuracy

Average of the magnitude of the errors

Linear score = each error has same weight

It does not indicates the direction of the error, just the magnitude

Median Absolute Deviation

$$MAD = median\{|f_i - o_i|\}$$

Attribute: measures accuracy

Median of the magnitude of the errors

Very robust

Extreme errors have no effect

Continuous scores: MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2$$

Attribute: measures accuracy

Average of the squares of the errors: it measures the magnitude of the error, weighted on the squares of the errors

it does not indicate the direction of the error

Quadratic rule, therefore large weight on large errors:

→ good if you wish to penalize large error

→ sensitive to large values (e.g. precipitation) and outliers; sensitive to large variance (high resolution models); encourage conservative forecasts (e.g. climatology)

Continuous scores: RMSE

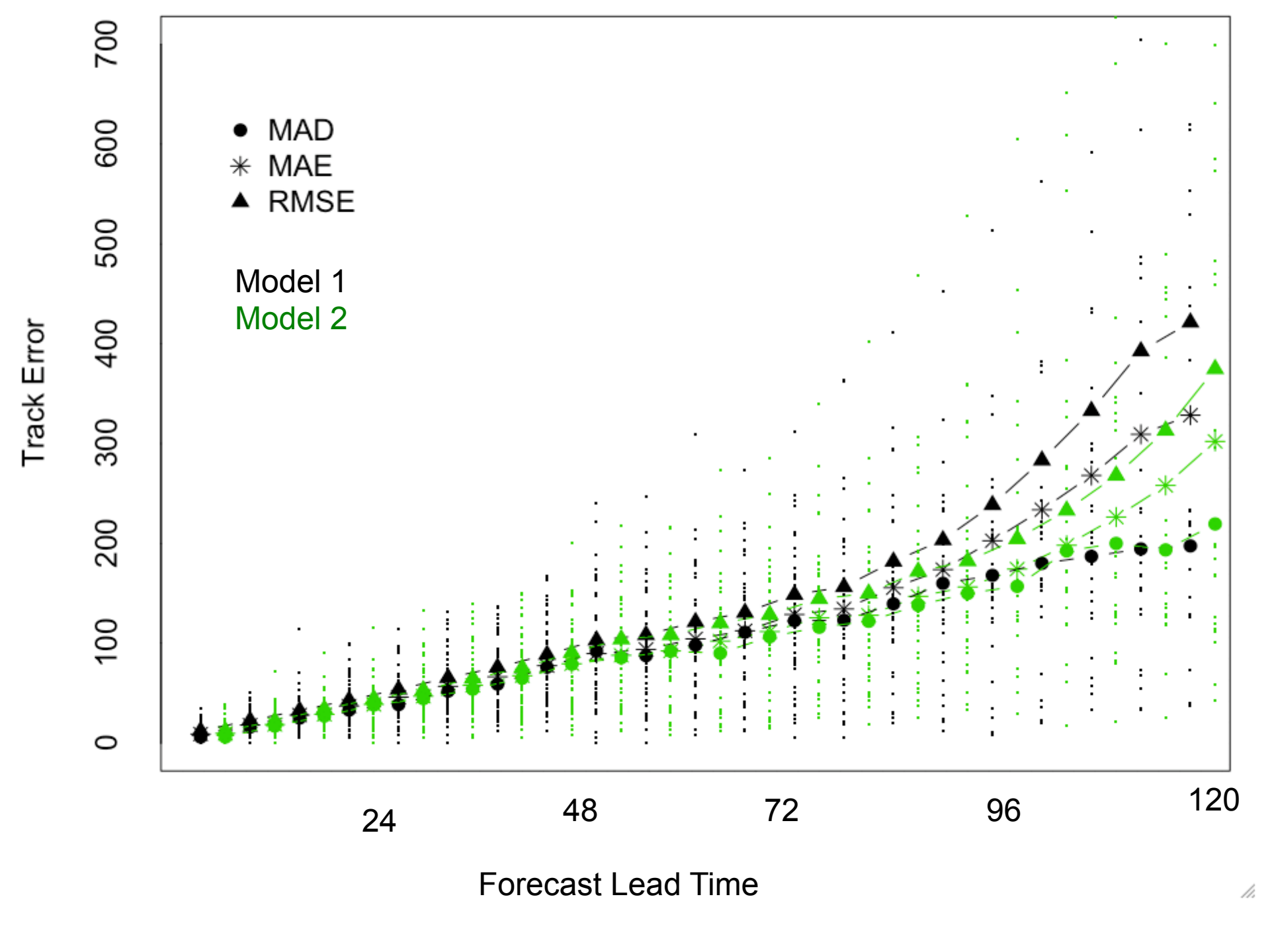
$$RMSE = \sqrt{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2$$

Attribute: measures accuracy

RMSE is the squared root of the MSE: measures the magnitude of the error retaining the variable unit (e.g. ^{o}C)

Similar properties of MSE: it does not indicate the direction the error; it is defined with a quadratic rule = sensitive to large values, etc.

NOTE: RMSE is always larger or equal than the MAE


MAD
MAE
RMSE
Model 1
Model 2
Track Error
0
100
200
300
400
500
600
700
24
48
72
96
120
Forecast Lead Time

Continuous scores: linear correlation

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = \frac{\mathrm{cov}(Y,X)}{s_Y s_X}$$

Attribute: measures association

Measures linear association between forecast and observation

Y and X rescaled (non-dimensional) covariance: ranges in [-1,1]

It is not sensitive to the bias

The correlation coefficient alone does not provide information on the inclination of the regression line (it says only is it is positively or negatively tilted); observation and forecast variances are needed; the slope coefficient of the regression line is given by $b = (s_X/s_Y)r_{XY}$

Not robust = better if data are normally distributed
Not resistant = sensitive to large values and outliers

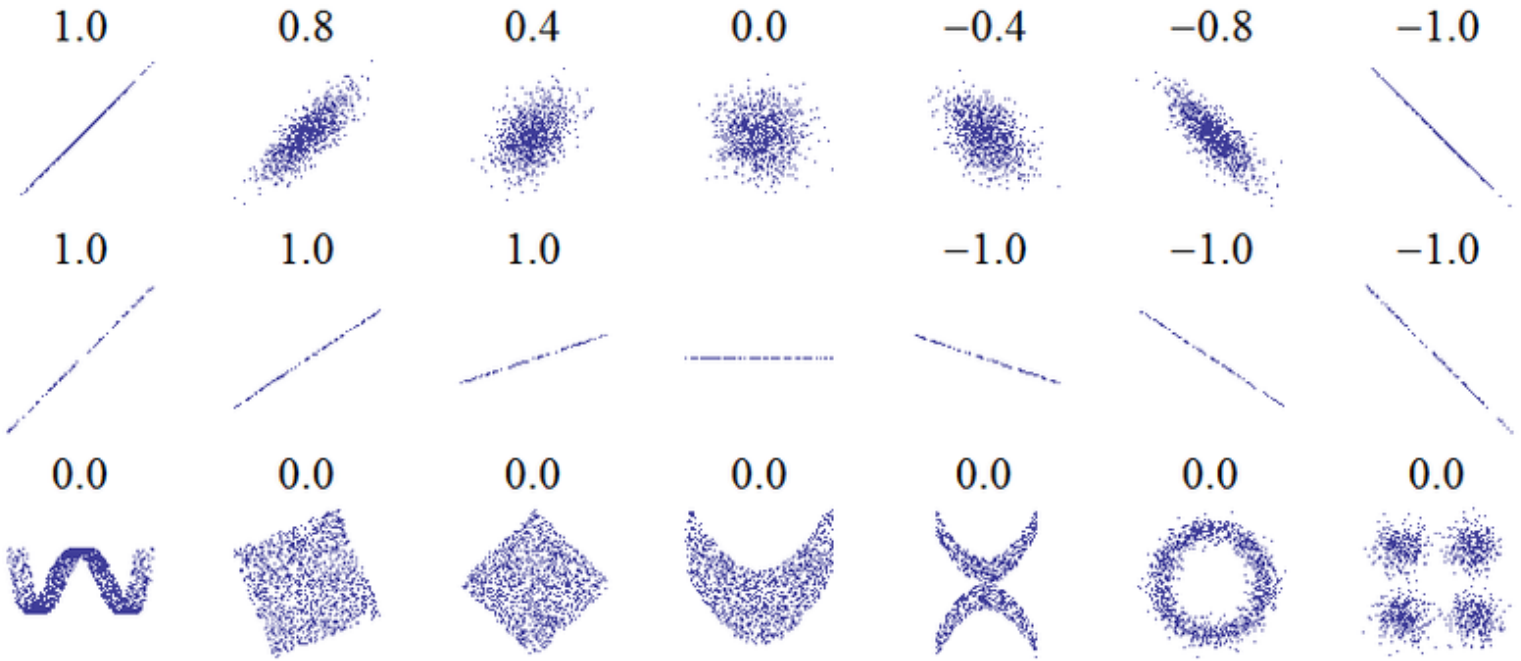
Scores for continuous forecasts

Simplest overall measure of performance:

Correlation coefficient

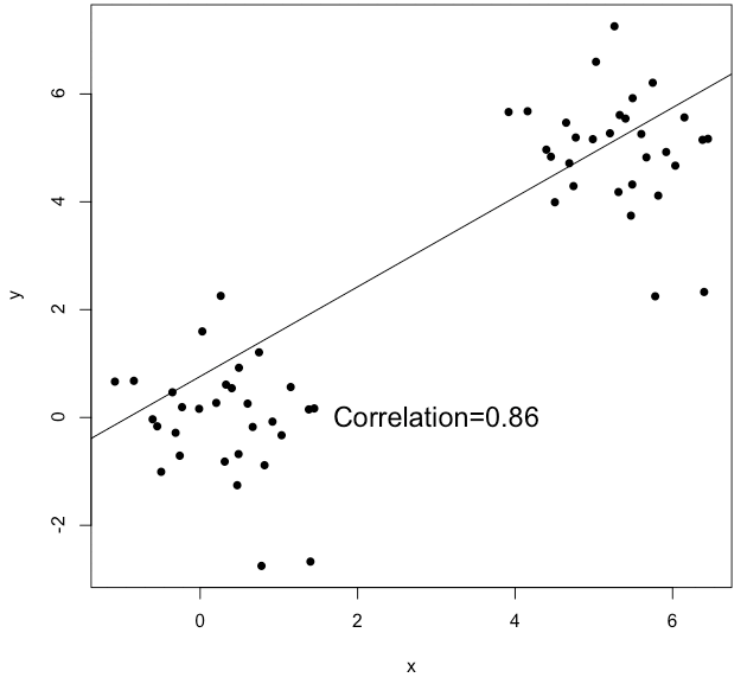
$$\rho_{fx} = \frac{Cov(f,x)}{\sqrt{Var(f)Var(x)}}$$

$$r_{fx} = \frac{\sum_{i=1}^{n}(f_i - \bar{f})(x_i - \bar{x})}{(n-1)s_f s_x}$$



Continuous scores: anomaly correlation

- Correlation calculated on anomaly.
- Anomaly is difference between what was forecast (observed) and climatology.
- Centered or uncentered versions.



MSE and bias correction

$$MSE = \left(\bar{f} - \bar{o}\right)^2 + s_f^2 + s_o^2 - 2s_f s_o r_{fo}$$

$$MSE = ME^2 + \mathrm{var}(f - o)$$

- MSE is the sum of the squared bias and the variance. So ↑ bias = ↑ MSE
- Bias and RMSE are ***not*** independent measures!
- *var(f – o)* is sometimes called *bias-corrected MSE*
- *Recommendation:* Report Bias (ME) and Bias-corrected MSE

Continuous skill scores: MAE skill score

$$SS_{MAE} = \frac{MAE - MAE_{ref}}{MAE_{perf} - MAE_{ref}} = 1 - \frac{MAE}{MAE_{ref}}$$

Attribute: measures skill

Skill score: measure the forecast accuracy with respect to the accuracy of a reference forecast: positive values = skill; negative values = no skill

Difference between the score and a reference forecast score, normalized by the score obtained for a perfect forecast minus the reference forecast score (for perfect forecasts MAE=0)

Reference forecasts:

- **persistence:** appropriate when time-correlation > 0.5
- **sample climatology:** information only a posteriori
- **actual climatology:** information a priori

Continuous skill scores: MSE skill score

$$SS_{MSE} = \frac{MSE - MSE_{ref}}{MSE_{perf} - MSE_{ref}} = 1 - \frac{MSE}{MSE_{ref}}$$

Attribute: measures skill

Same definition and properties as the MAE skill score: measure accuracy with respect to reference forecast, positive values = skill; negative values = no skill

Sensitive to sample size (for stability) and sample climatology (e.g. extremes): needs large samples

Reduction of Variance: MSE skill score with respect to climatology.
If sample climatology is considered:

$$Y = \bar{X}; \quad MSE_{cli} = s_X^2 \quad \text{and} \quad RV = 1 - \frac{MSE}{s_X^2} = r_{XY}^2 - \left(r_{XY} - \frac{s_Y}{s_X} \right)^2 - \left(\frac{\bar{Y} - \bar{X}}{s_X} \right)^2$$

linear correlation

bias

reliability: regression line slope coeff $b=(s_X/s_Y)r_{XY}$

Continuous Scores of **Ranks**

Problem: Continuous scores sensitive to large values or non robust.

Solution: Use the **ranks** of the variable, rather than its actual values.

Temp °C	27.4	21.7	24.2	23.1	19.8	25.5	24.6	22.3
rank	8	2	5	4	1	7	6	3

The value-to-rank transformation:

- diminish effects due to large values
- transform distribution to a Uniform distribution
- remove bias

Rank correlation is the most common.

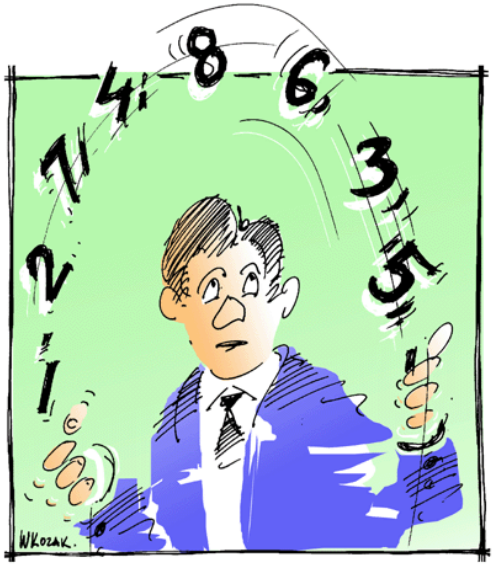
Conclusions

- Verification information can help you better understand and improve your forecasts.
- This session has only begun to cover basic verification topics.
- Additional topics and information are available.
- Advanced techniques may be needed to evaluate and utilize forecasts effectively.
  - Confidence intervals
  - Spatial and diagnostic methods
  - Ensemble and probabilistic methods

Software:

MET (Model Evaluation Tools) software.
www.dtcenter.org/met/users

R Verification package.
www.cran.r-project.org/web/packages/verification.index.html



References:

Jolliffe and Stephenson (2003): Forecast Verification: a practitioner's guide, Wiley & Sons, 240 pp.

Wilks (2011): Statistical Methods in Atmospheric Science, Academic press, 467 pp.

Stanski, Burrows, Wilson (1989) Survey of Common Verification Methods in Meteorology

http://www.eumetcal.org.uk/eumetcal/verification/www/english/courses/msgcrs/index.htm

http://www.cawcr.gov.au/projects/verification/verif_web_page.html