Verification of Continuous Forecasts

Presented by
Barbara Brown

Including contributions by
Tressa Fowler, Barbara Casati, Laurence Wilson, and others

DTC
Developmental Testbed Center

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

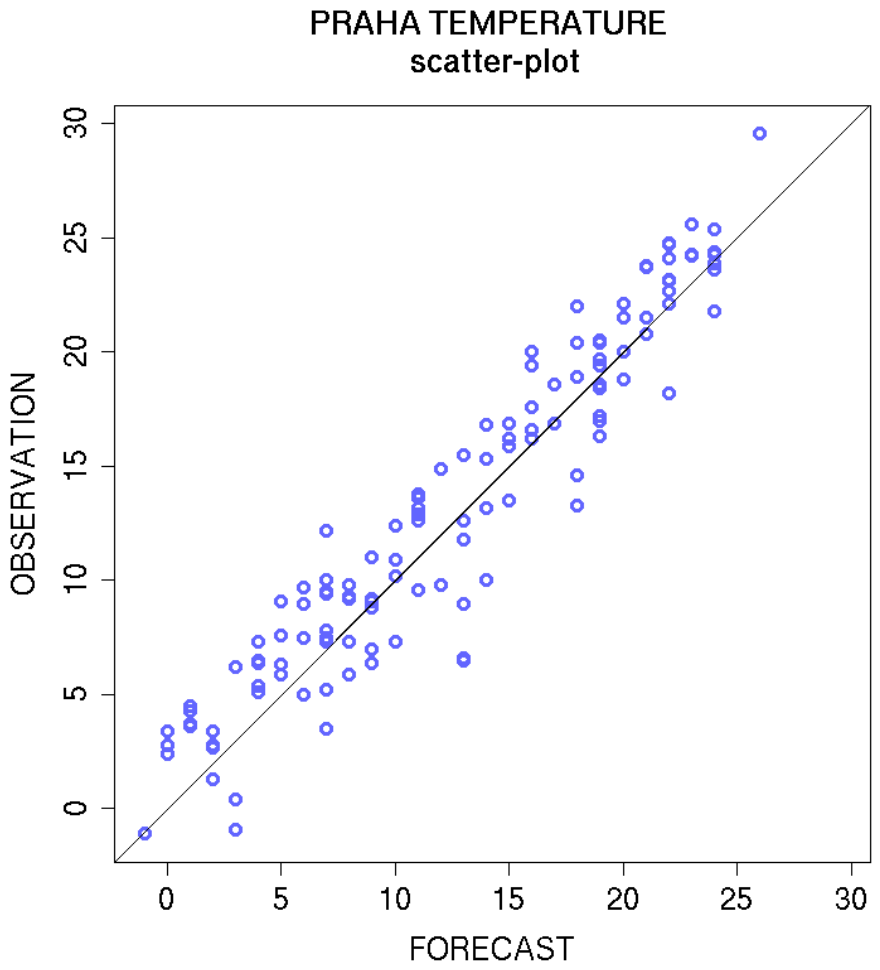
- Statistics
  - Bias
  - Error statistics
  - Robustness
  - Comparisons

Exploratory methods: joint distribution

Scatter-plot: plot of observation versus forecast values

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

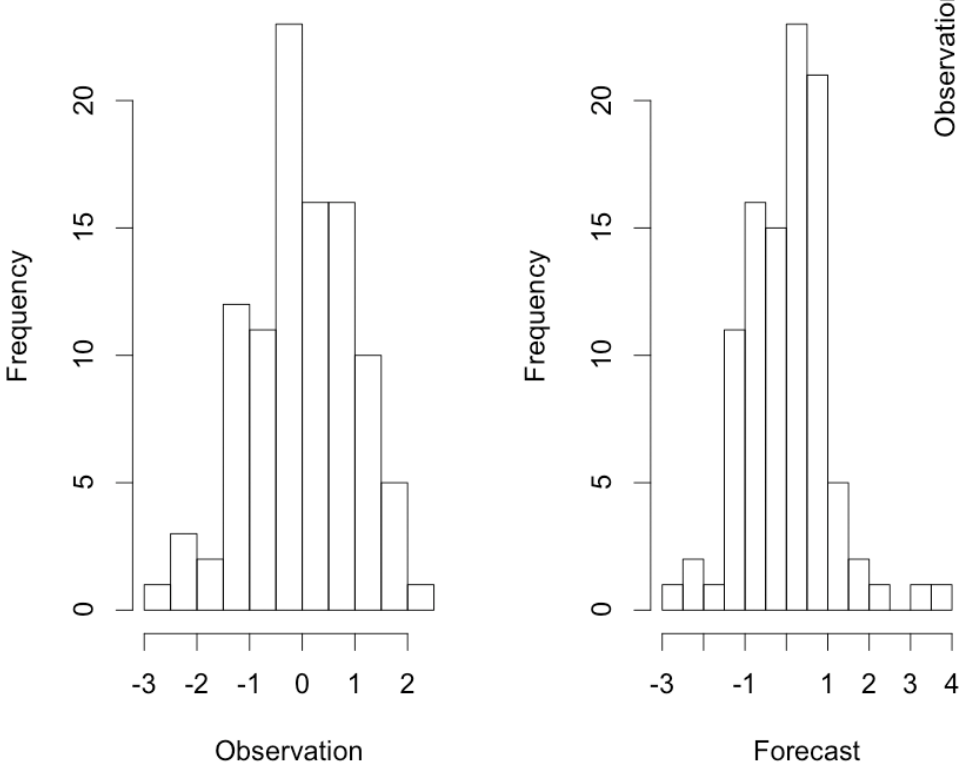
Provides information on: bias, outliers, error magnitude, linear association, peculiar behaviours in extremes, misses and false alarms (link to contingency table)

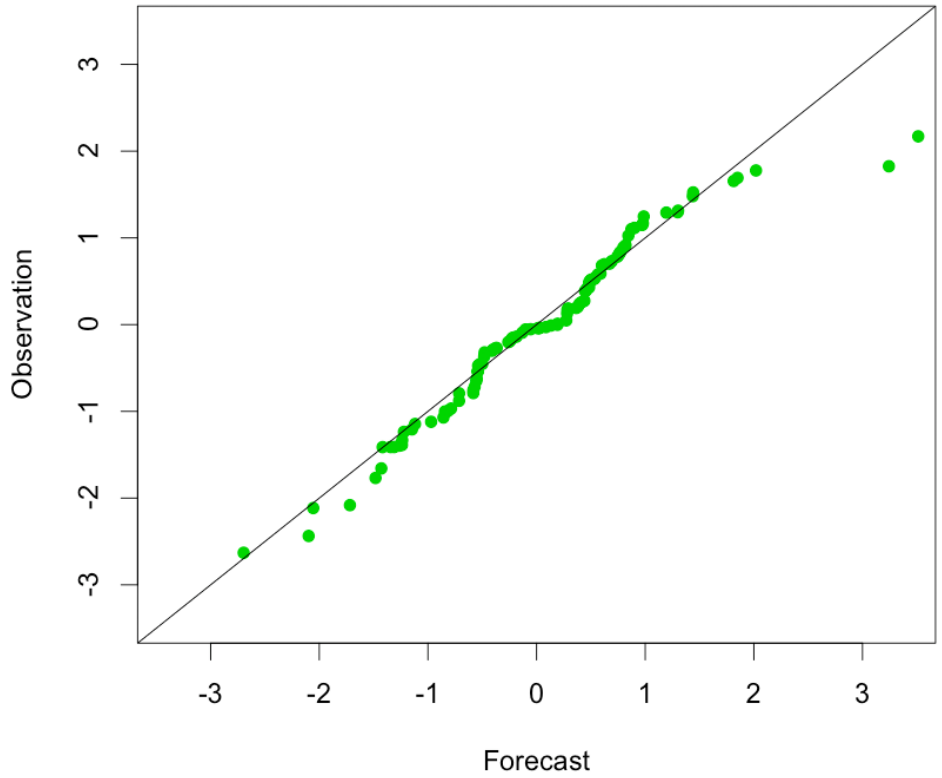


Exploratory methods: marginal distribution

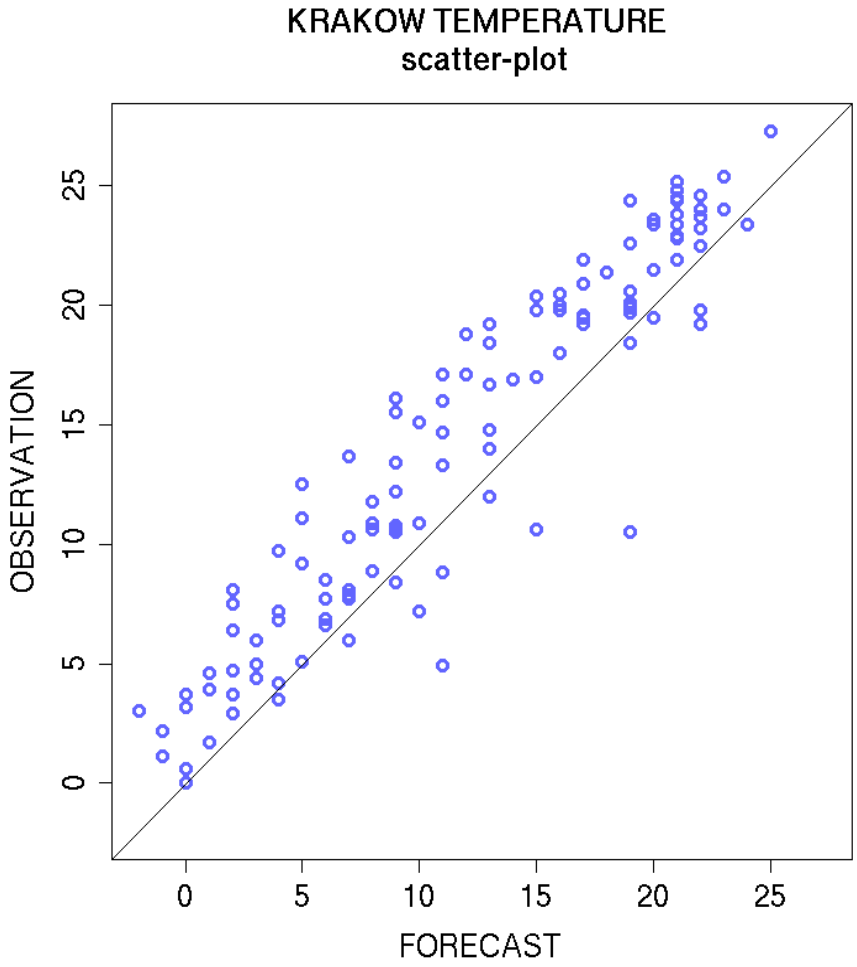
Quantile-quantile plots:

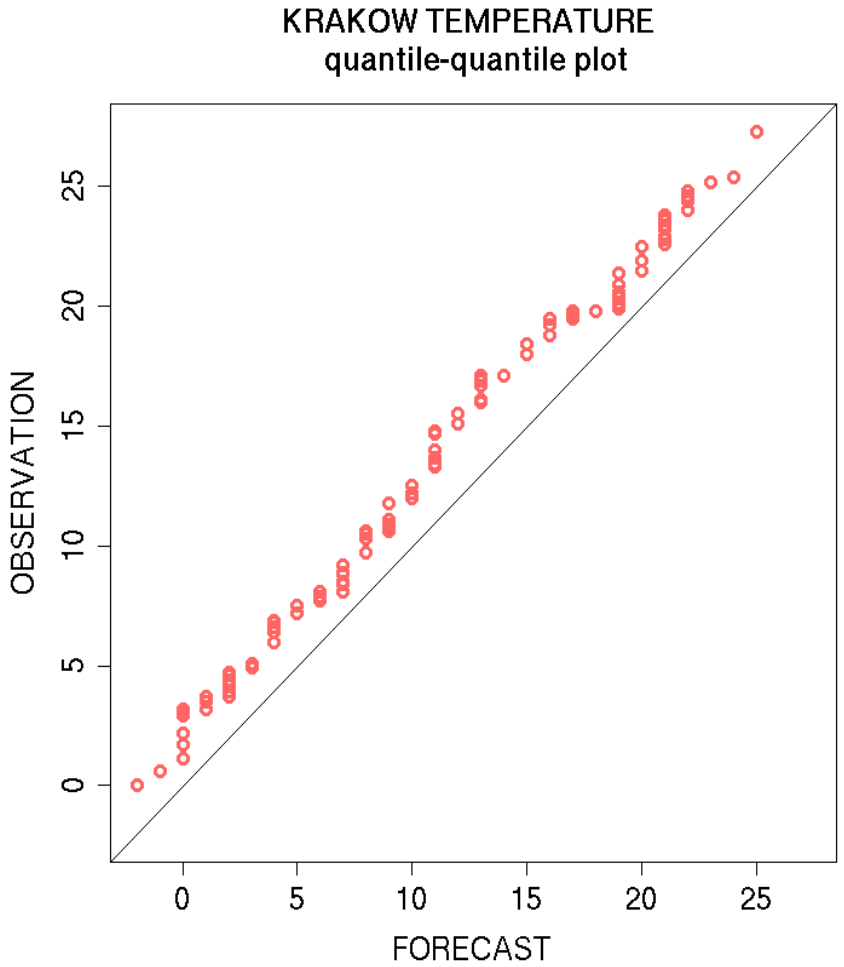
OBS quantile versus the corresponding FRCS quantile





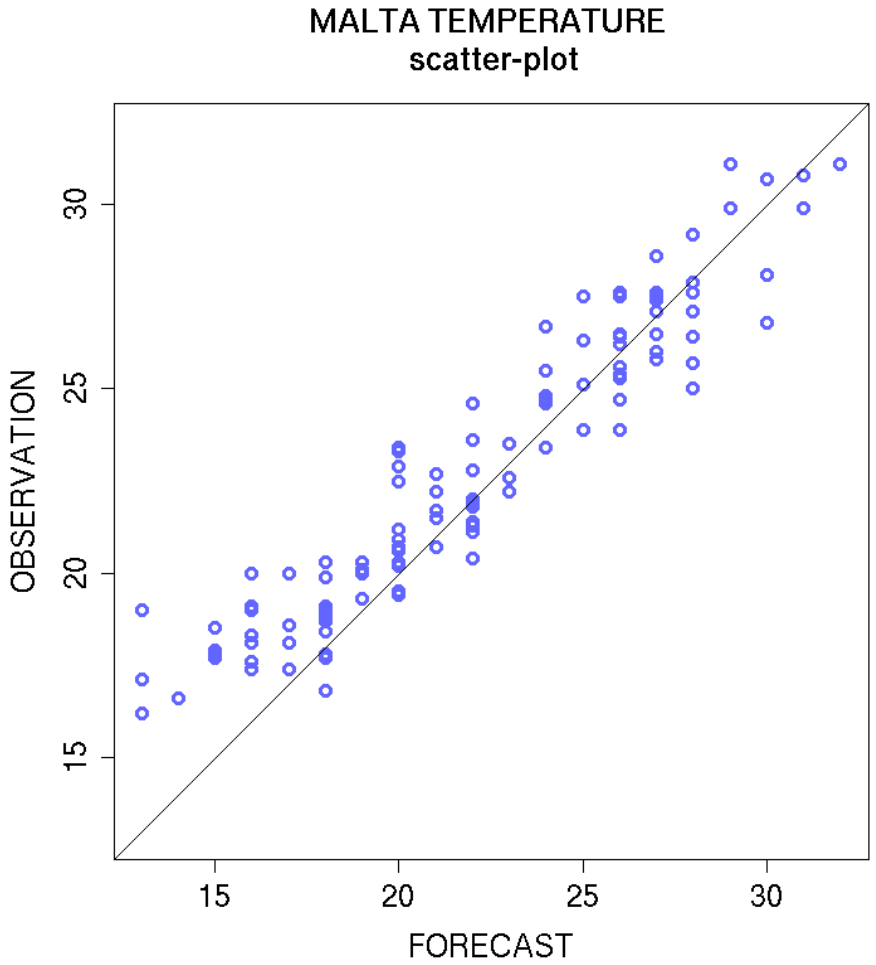
Scatter-plot and qq-plot: example 1
Q: is there any bias? Positive (over-forecast) or negative (under-forecast)?

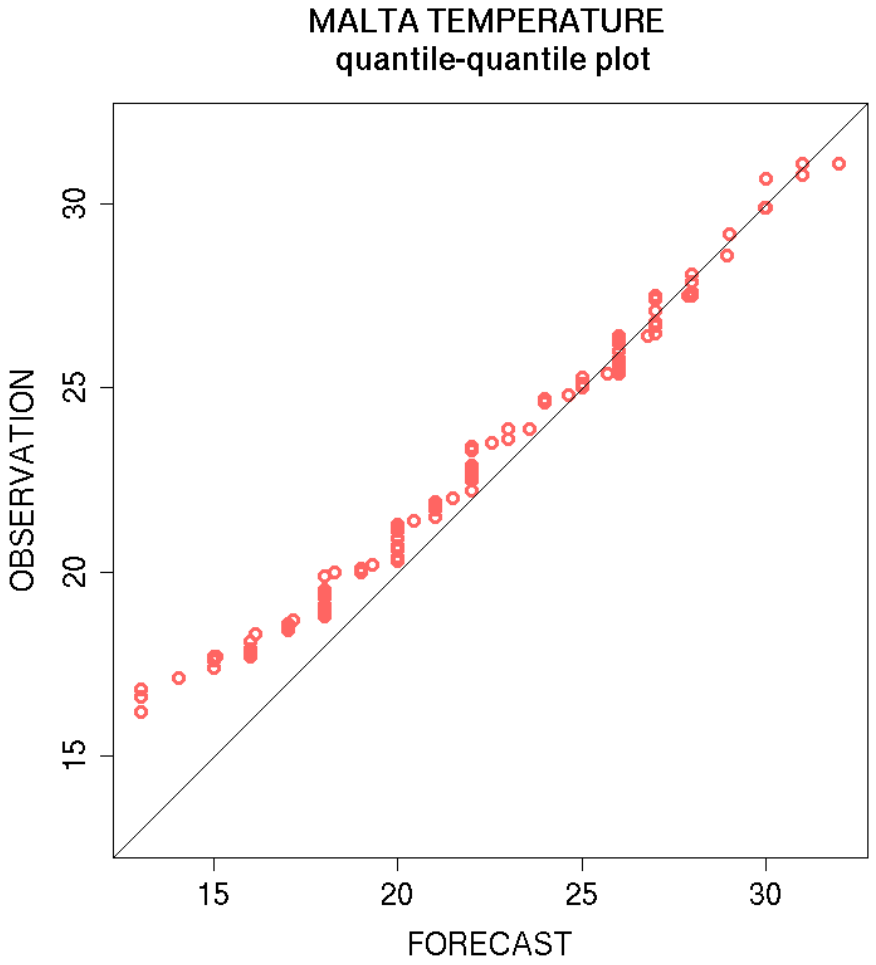




Scatter-plot and qq-plot: example 2

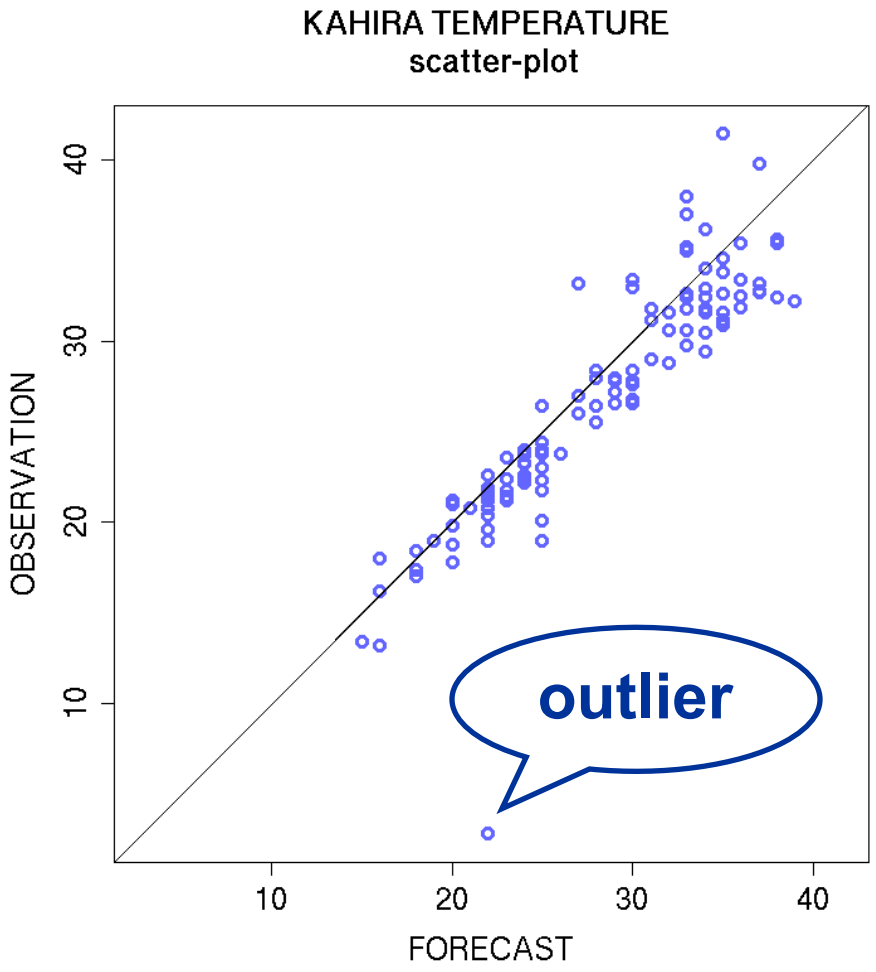
Describe the peculiar behaviour of low temperatures

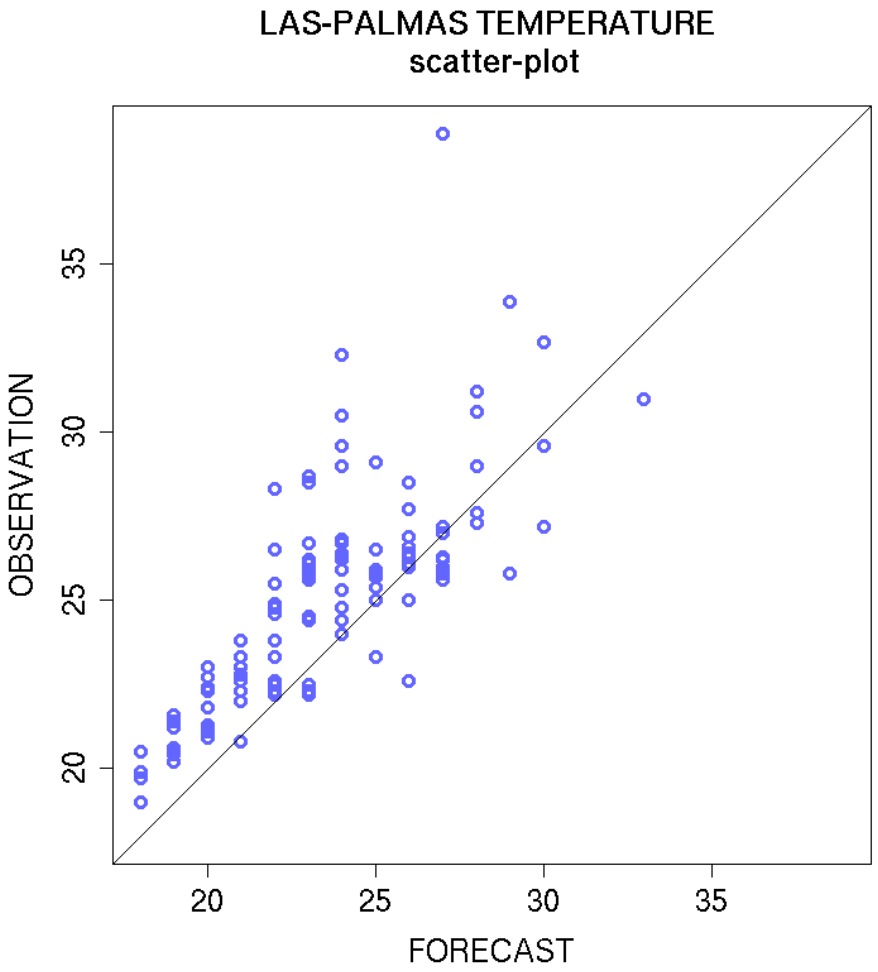




Scatter-plot: example 3

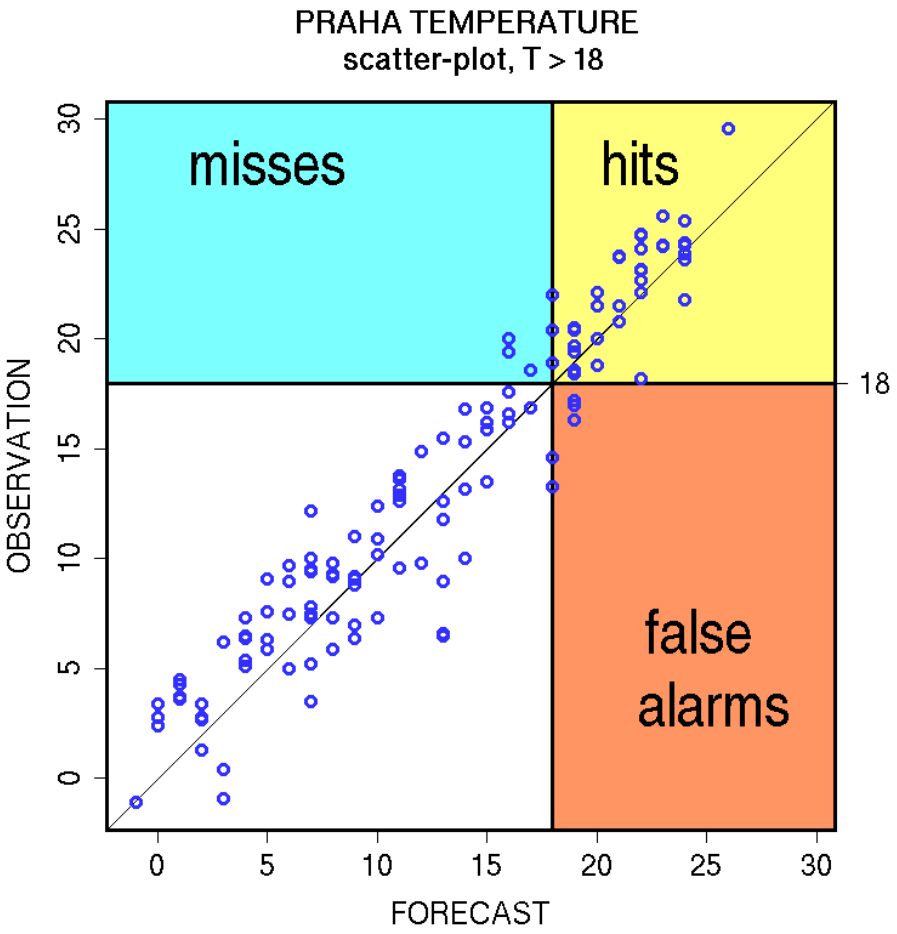
Describe how the error varies as the temperatures grow



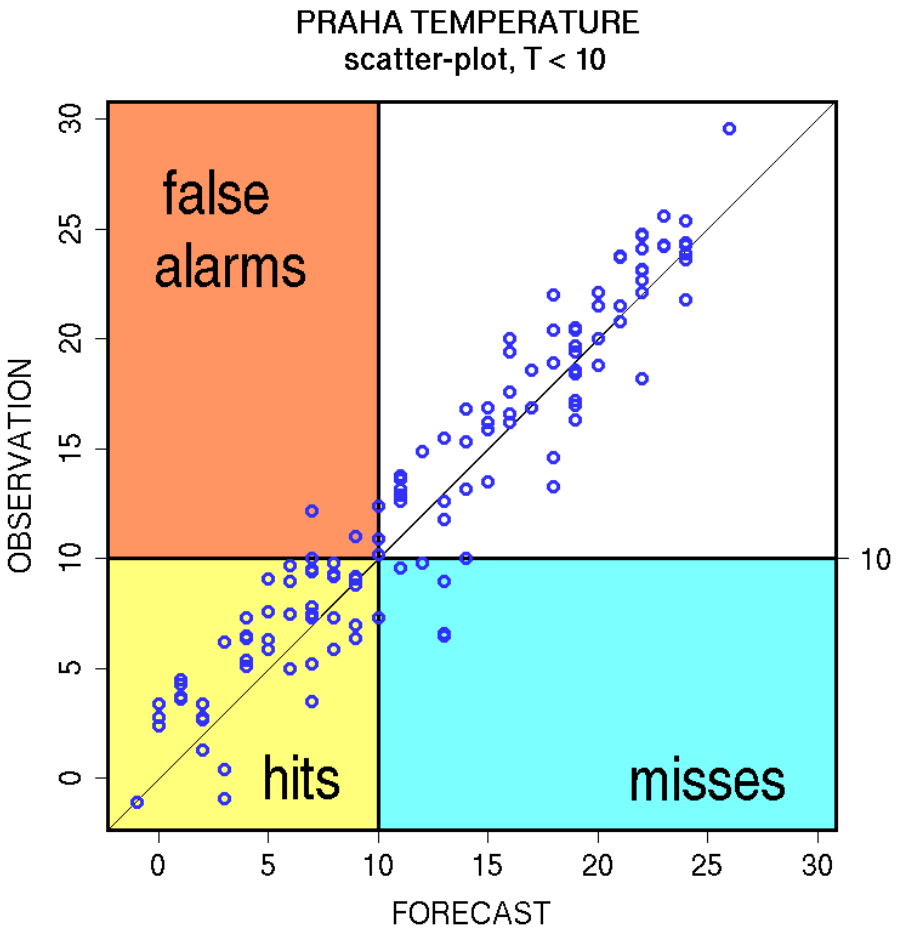


Scatter-plot and Contingency Table

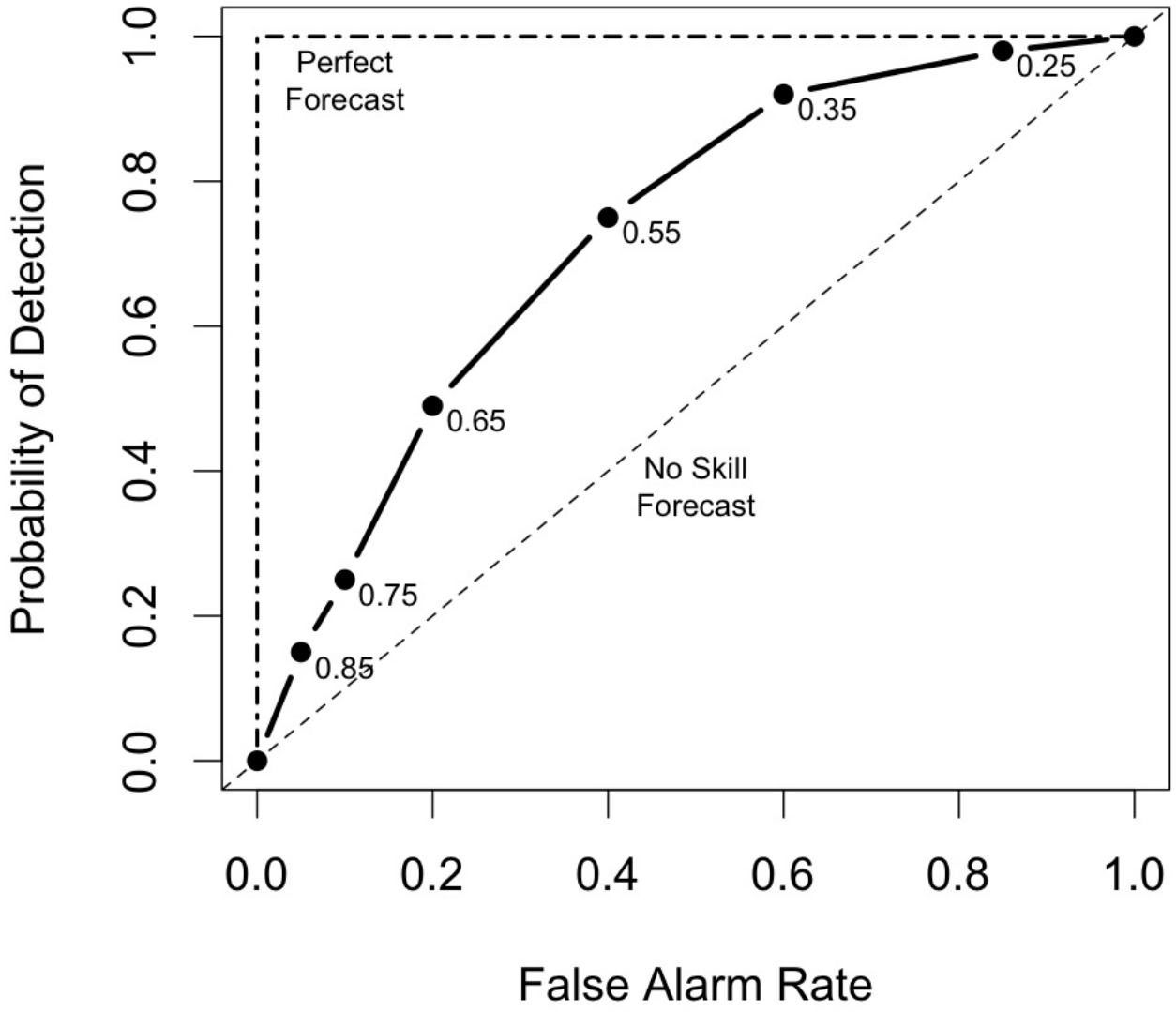
Does the forecast detect correctly temperatures above 18 degrees ?



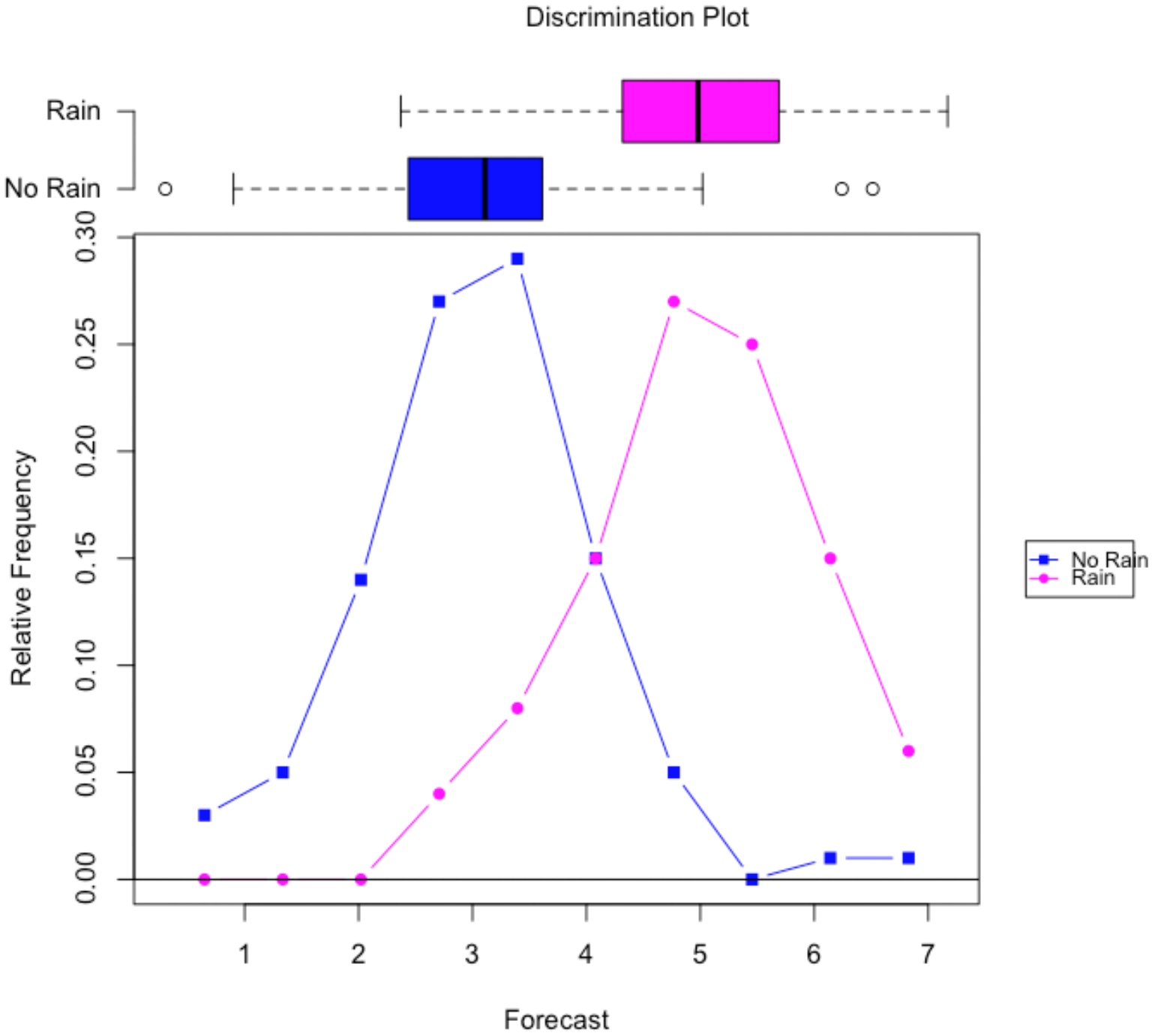
Does the forecast detect correctly temperatures below 10 degrees ?



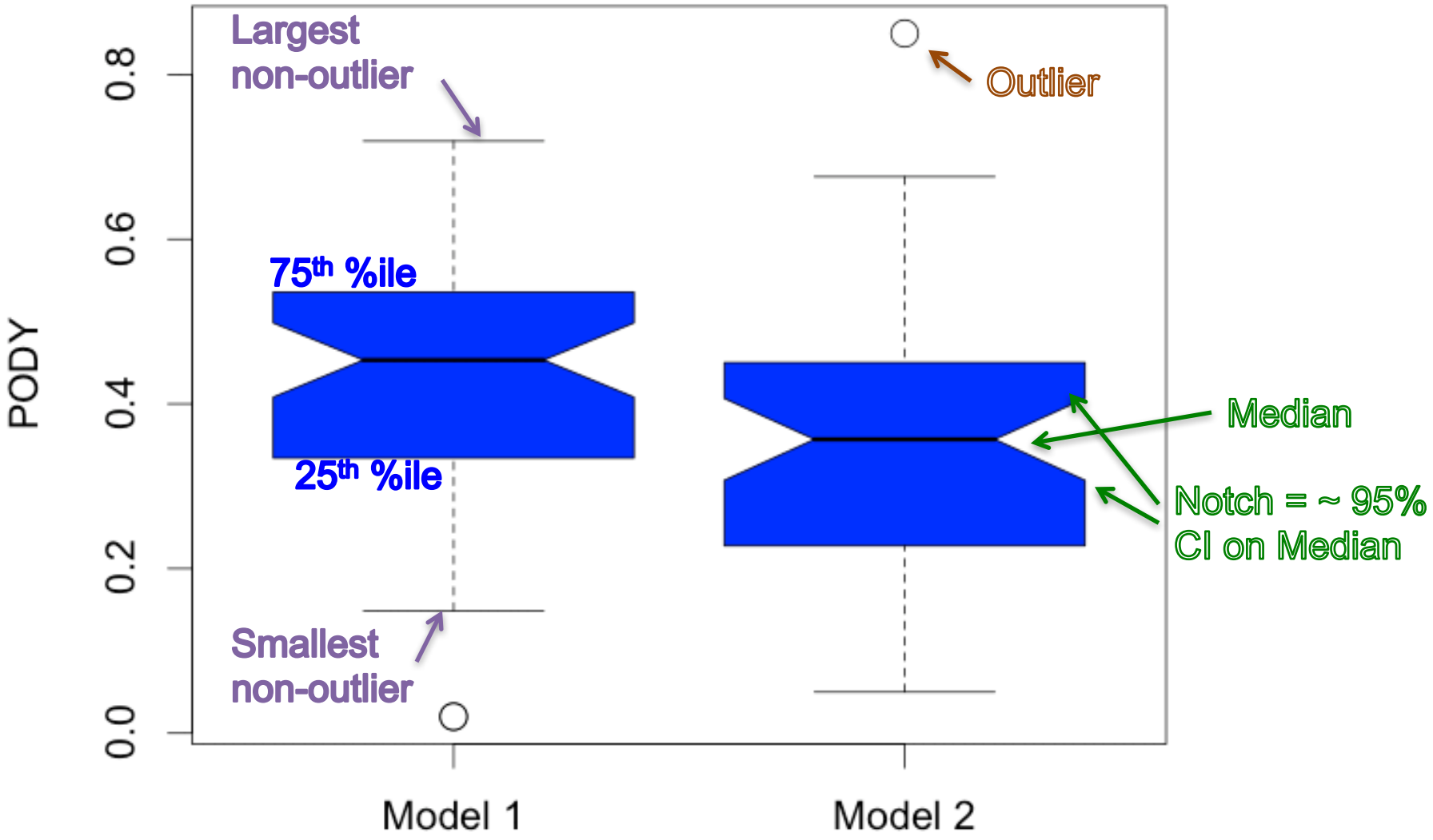
Example Receiver Operating Characteristic Plot



Create with points from PRC line type


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:***

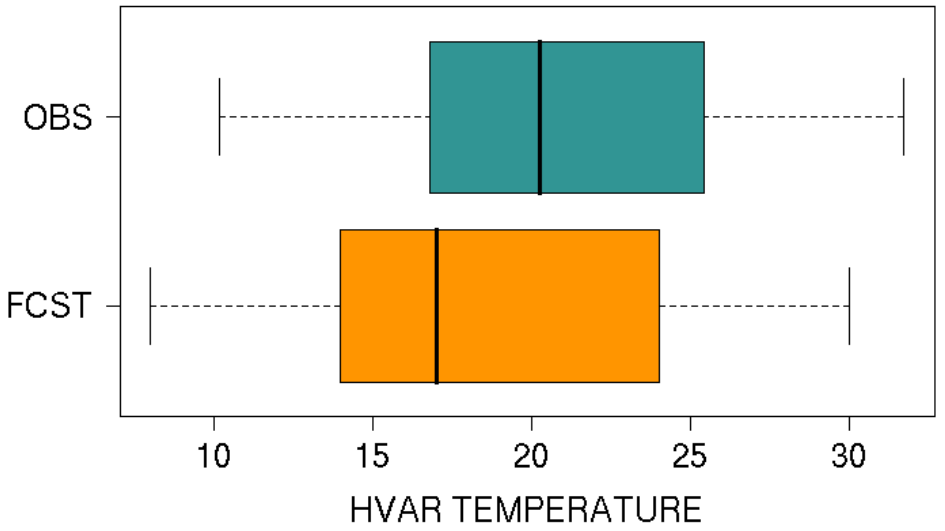
$$\text{mean} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

- ***Spread:***

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

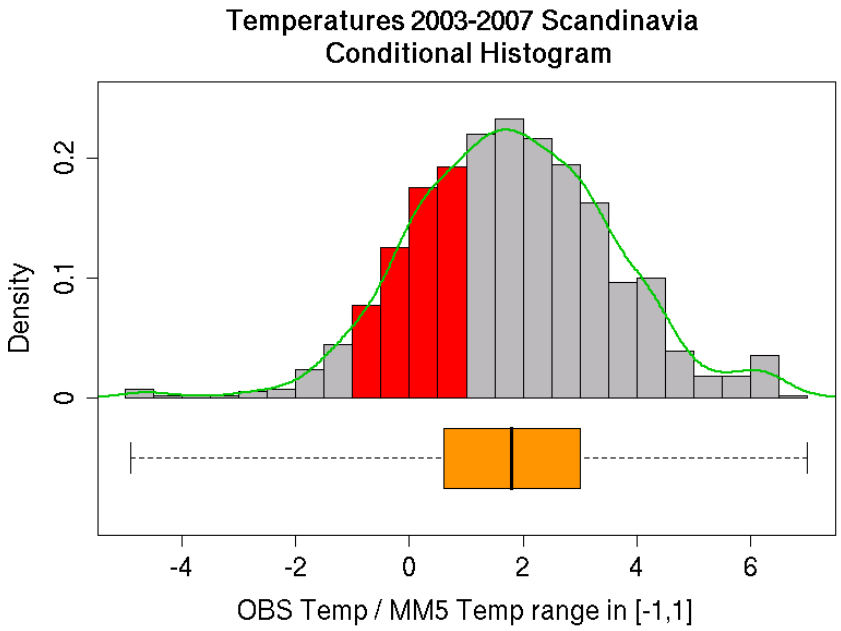
$$\text{Inter Quartile Range} = \text{IQR} = q_{0.75} - q_{0.25}$$

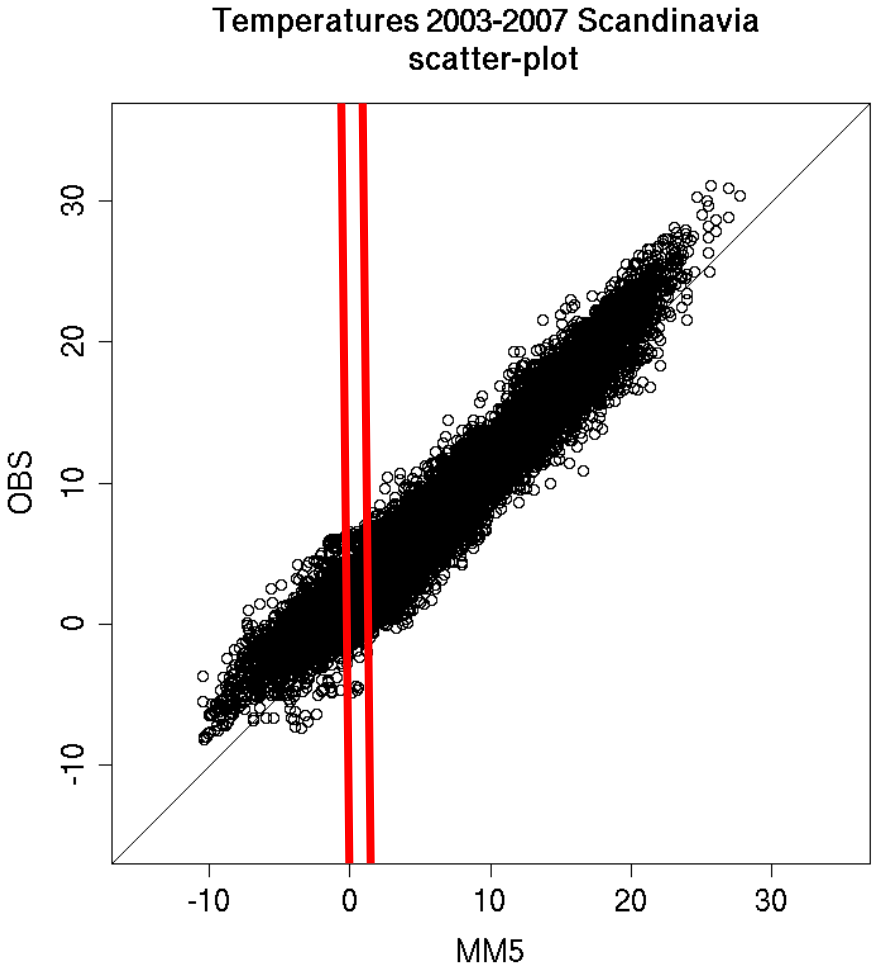


	MEAN	MEDIAN	STDEV	IQR
OBS	20.71	20.25	5.18	8.52
FRCS	18.62	17.00	5.99	9.75

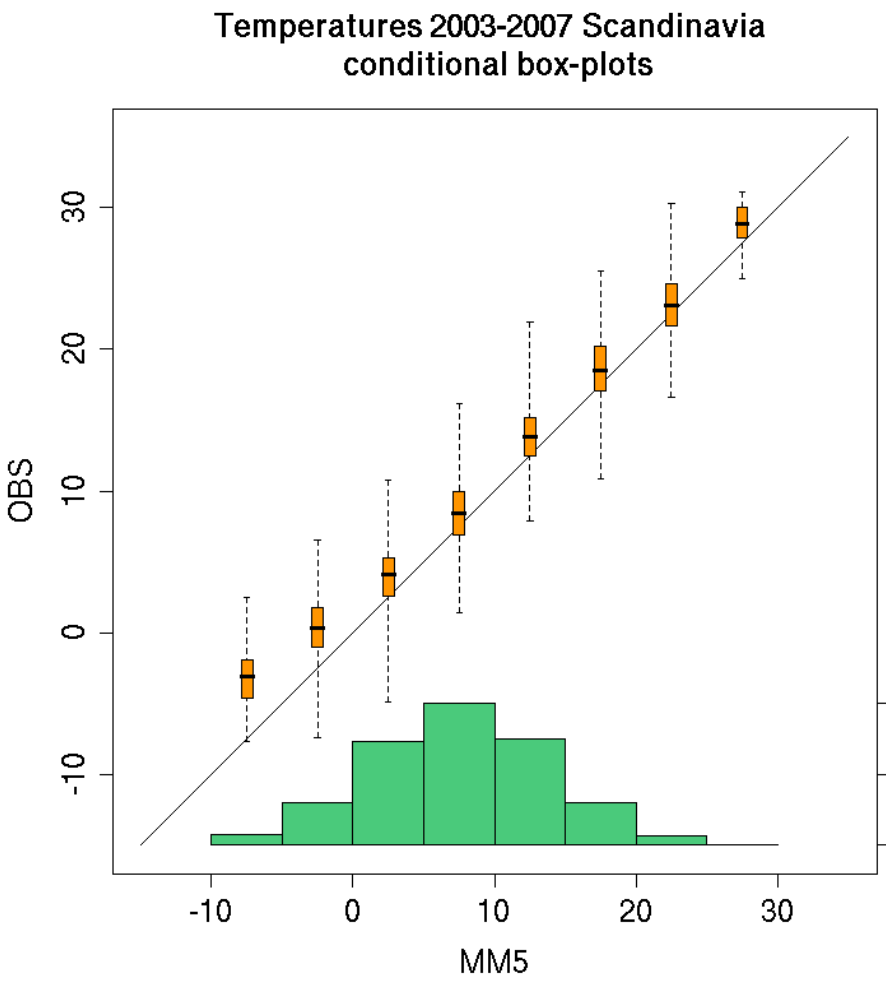
Exploratory methods: conditional distributions

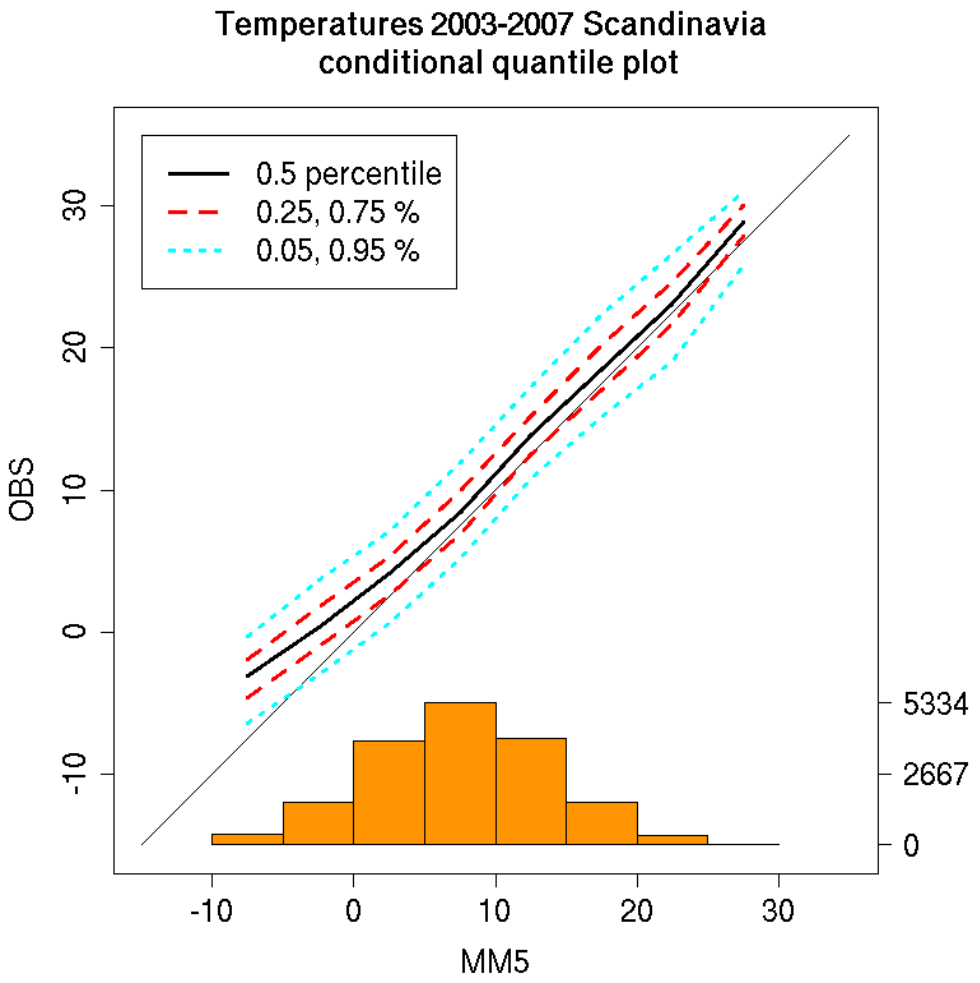
Conditional histogram and conditional box-plot





Exploratory methods: conditional quantile 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

  Indicates the *average direction of error*:
  - positive bias indicates over-forecast,
  - negative bias indicates under-forecast
- 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 ?

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
- Does not indicates the *direction* of the error, just the *magnitude*

Median Absolute Deviation

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

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*:
  - measures the *magnitude* of the error, weighted on the squares of the errors
  - does not indicate the *direction* of the error

- Quadratic rule (large weight on large errors)
  - ❑ good if you wish to penalize large errors
  - ❑ sensitive to large values (e.g., precipitation) and outliers;
  - ❑ sensitive to large variance (high resolution models);
  - ❑ encourages conservative forecasts (e.g. climatology)

Continuous scores: RMSE

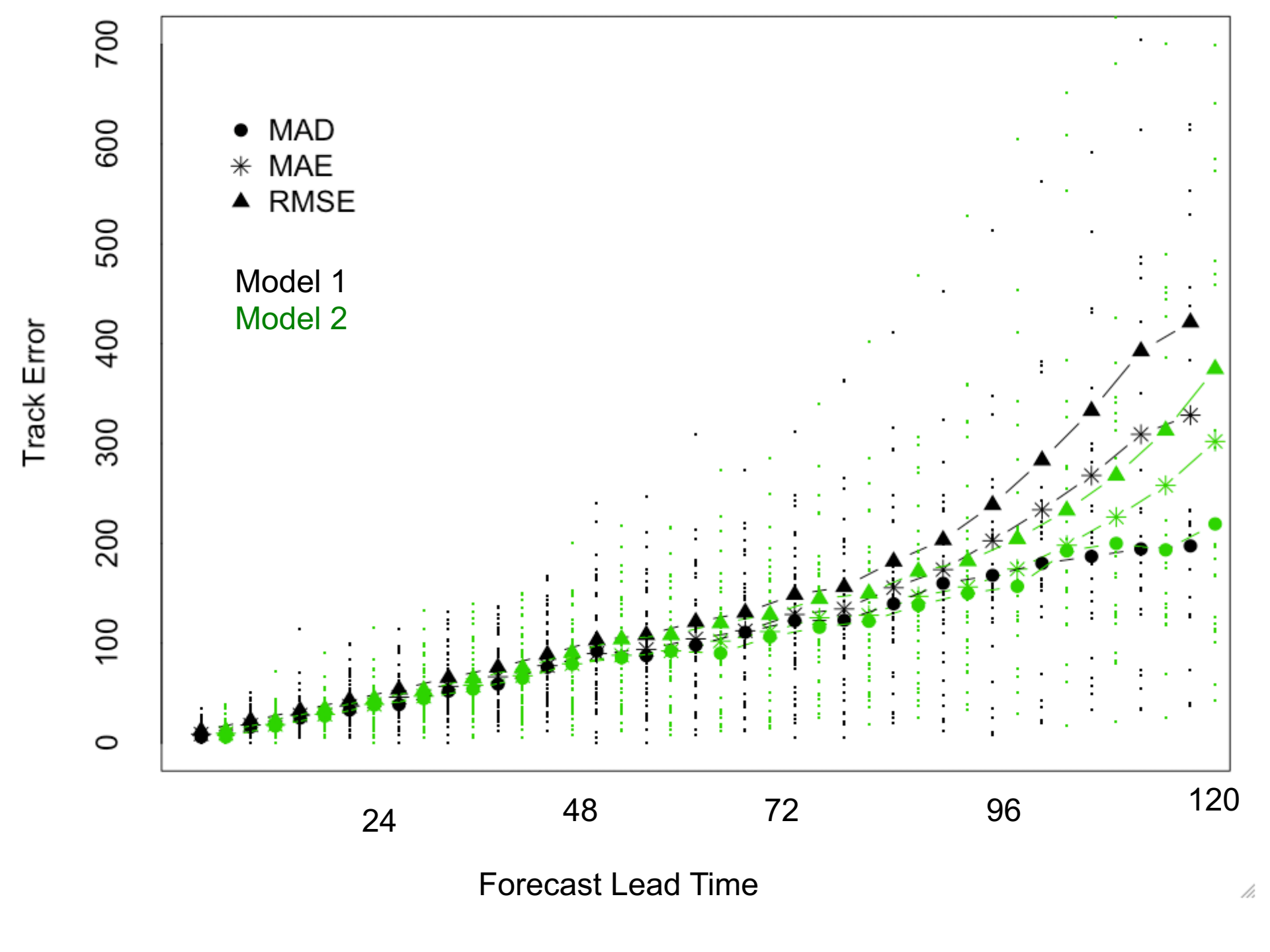
$$RMSE = \sqrt{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2$$

Attribute: measures accuracy

- Square root of 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

$$r_{FO} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - \bar{f}\right)\left(o_i - \bar{o}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - \bar{f}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\left(o_i - \bar{o}\right)^2}} = \frac{\mathrm{cov}(F,O)}{s_F s_O}$$

Attribute: measures association

- Measures *linear association* between forecast and observation
- F and O rescaled (non-dimensional) covariance: range is [-1,1]
- Not sensitive to the *bias*
- Alone does not provide information on the slope of the regression line (it says only is it is positively or negatively tilted
- *Not robust* = better if data are normally distributed
- *Not resistant* = sensitive to large values and outliers

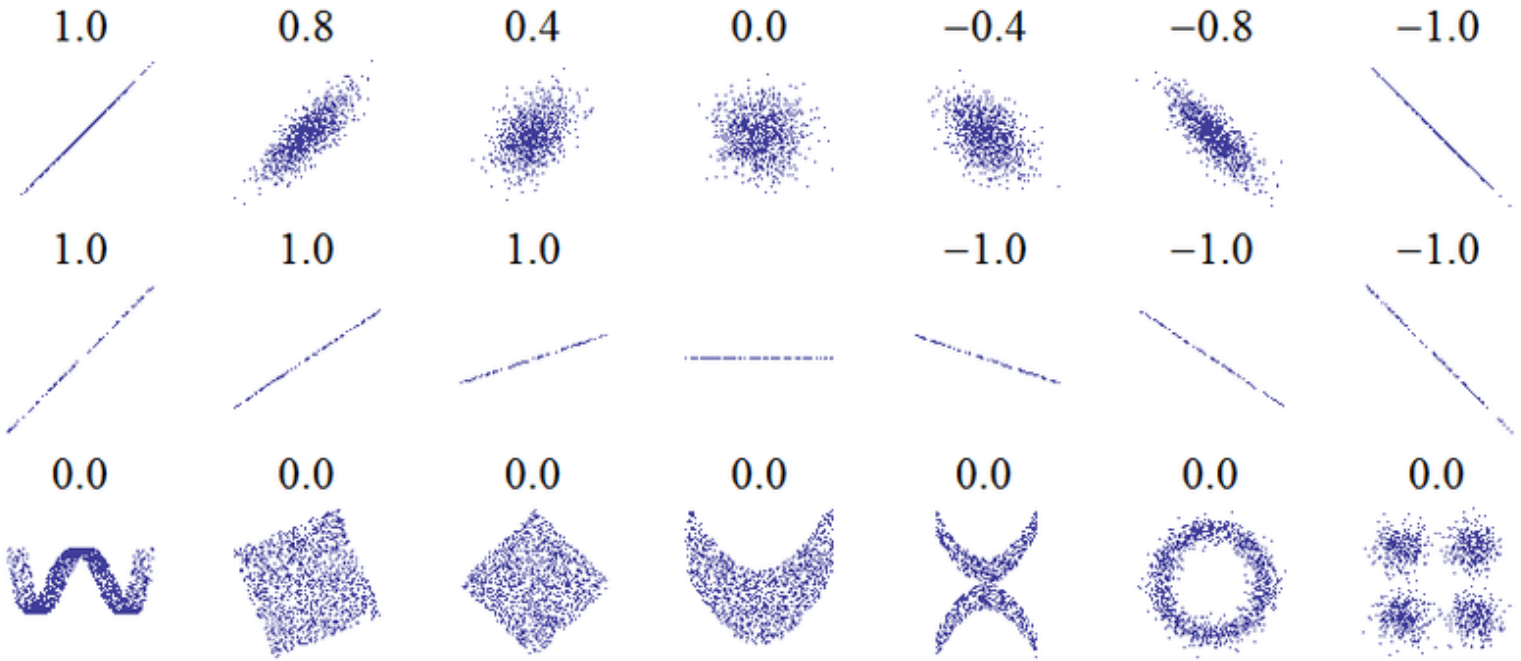
Scores for continuous forecasts

Simplest overall measure of performance:

Correlation coefficient

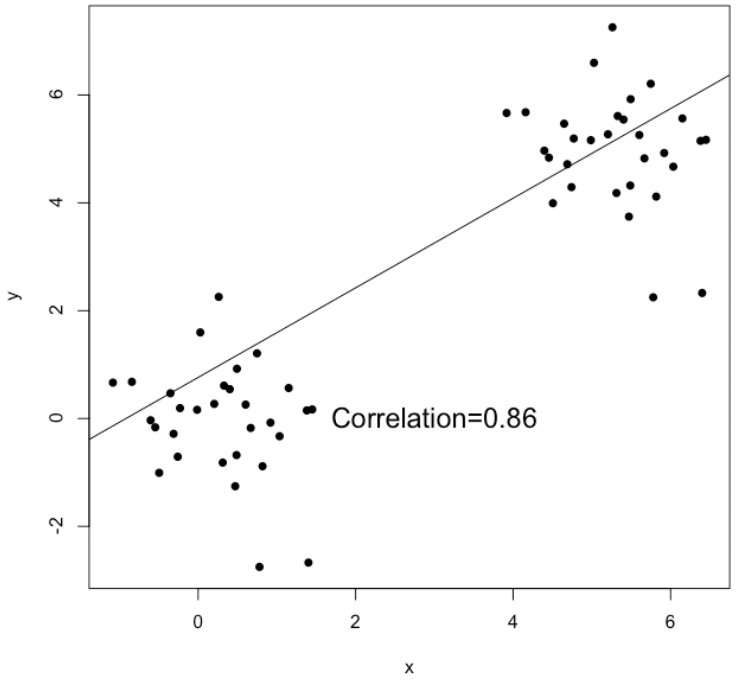
$$\rho_{fx} = \frac{Cov(f,x)}{\sqrt{Var(f)Var(x)}}$$

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



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

Continuous skill scores: MAE skill score

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

Attribute: measures skill

- A *skill score* measures the forecast accuracy relative to the accuracy of a *reference forecast*
  - positive values => forecast is skillful
  - 0 value => no skill
  - negative values => negative skill
- Reference forecasts:
  - *Persistence*: Appropriate when time-correlation > 0.5
  - *Sample climatology*: Based on *a posteriori* information
  - *Actual climatology*: Based on *a priori* information

Continuous skill scores: MSE skill score

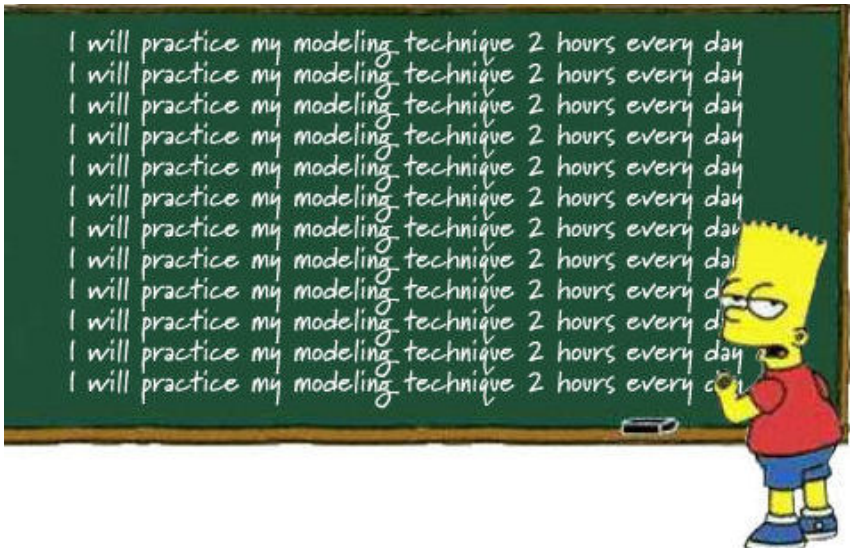
$$SS_{MSE} = \frac{MSE - MSE_{ref}}{MSE_{perf} - MSE_{ref}} = 1 - \frac{MSE}{MSE_{ref}}$$

Attribute: measures skill

- Same definition and properties as the MAE skill score
  - ❖ measure accuracy with respect to reference forecast,
  - ❖ positive values = positive skill;
  - ❖ negative values = negative skill;
  - ❖ 0 value = no skill
- Sensitive to sample size (for stability) and sample climatology (e.g., extremes): needs large samples

Continuous skill scores: good practice rules

- Use same climatology for the comparison of different models.
- In general, using **sample climatology** as a reference leads to a worse skill score than **long-term climatology**
  - *Always ask which climatology is used to evaluate the skill.*



Continuous skill scores: good practice rules

- If the climatology is calculated by (a) combining data from many different stations and times of the year, the skill score will be *better* than if (b) **a different climatology for each station and month of the year are used**.
  - In case (a) the model gets credit for correctly forecasting seasonal trends and climatologies at specific locations.
  - In case (b) the specific topographic effects and long-term trends are removed and the forecasts' discriminating capability is better evaluated.

  => It's important to choose the appropriate climatology to meet your verification purposes.

- *Persistence forecast*: use same time of the day to avoid diurnal cycle effects.

Continuous Scores of **Ranks**

Problem: Continuous scores can be sensitive to large values or not 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:

- diminishes effects due to large values
- transforms distribution to a Uniform distribution
- removes bias

Rank correlation is the most common.

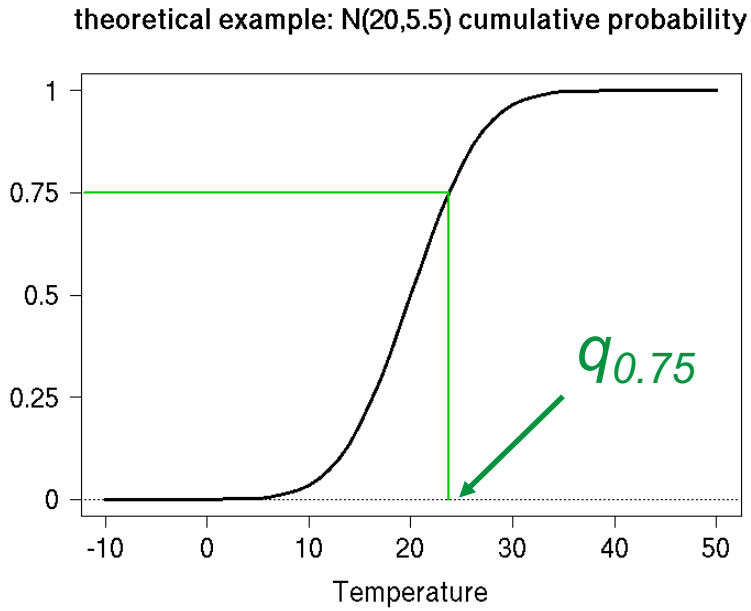
Linear Error in Probability Space

$$LEPS = \frac{1}{n}\sum_{i=1}^{n}\left|F_X(y_i) - F_X(x_i)\right|$$

The LEPS is a MAE evaluated by using the cumulative frequencies of the observation

Errors in the tail of the distribution are penalized less than errors in the centre of the distribution

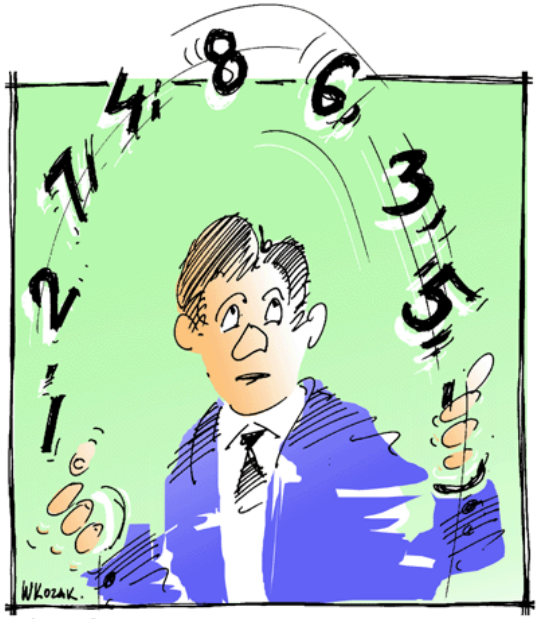
MAE and LEPS are minimized by the median correction



Summary

- Start with *exploratory graphics* – scatterplots, box plots, etc. – to depict marginal, conditional, joint distributions
- Many scores are available for evaluation of *continuous forecasts* – some are more useful than others
- Select scores that provide *meaningful answers* to the questions of interest!
  - Focus on Extremes? Bias? Average performance?
- *Decompose* scores where possible (e.g., MSE into Bias and squared errors
- *Skill scores* provide useful evaluation of forecast improvements
  - Use care in selecting standard of comparison (e.g., persistence, long-term climatology, sample climatology)

Thank you!



References:

Jolliffe and Stephenson (2012): Forecast Verification: a practitioner's guide, Wiley & Sons, 240 pp.

Wilks (2011): Statistical Methods in Atmospheric Science, Elsevier, 704 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

WMO Verification working group forecast verification web page, http://www.cawcr.gov.au/projects/verification/