Verification of forecasts of continuous variables

Manfred Dorninger
University of Vienna
Vienna, Austria
manfred.dorninger@univie.ac.at

Thanks to: B. Brown, M. Göber, B. Casati

7th Verification Tutorial Course, Berlin, 3-6 May, 2017

Types of forecasts, observations

- Continuous
  - Ex: Temperature, Rainfall amount, Humidity, Wind speed

- Categorical
  - Dichotomous (e.g., Rain vs. no rain, freezing or no freezing)
  - Multi-category (e.g., Cloud amount, precipitation type)
  - May result from *subsetting* continuous variables into categories
    - *Ex: Temperature categories of 0-10, 11-20, 21-30, etc.*

- Categorical approaches are often used when we want to truly "verify" something: i.e., was the forecast right or wrong?

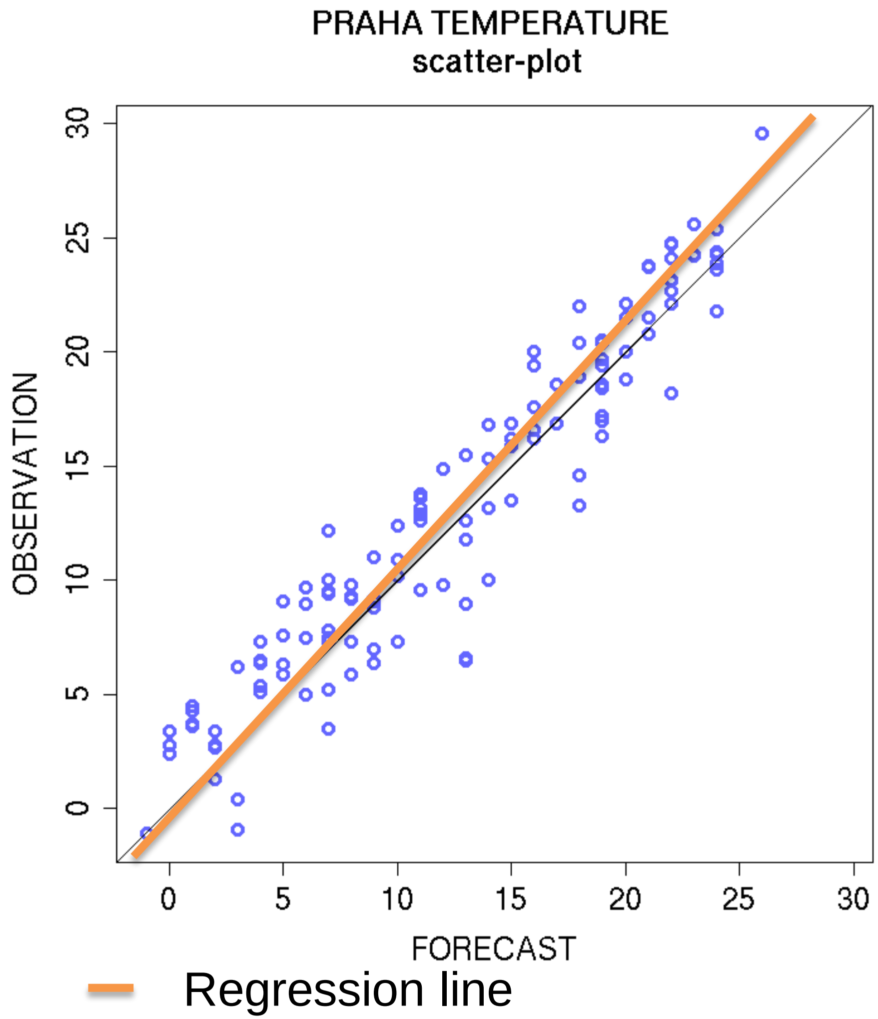
- Continuous approaches are often used when we want to know "how" they were wrong

Exploratory methods: joint distribution

Scatter-plot: plot of observation versus forecast values

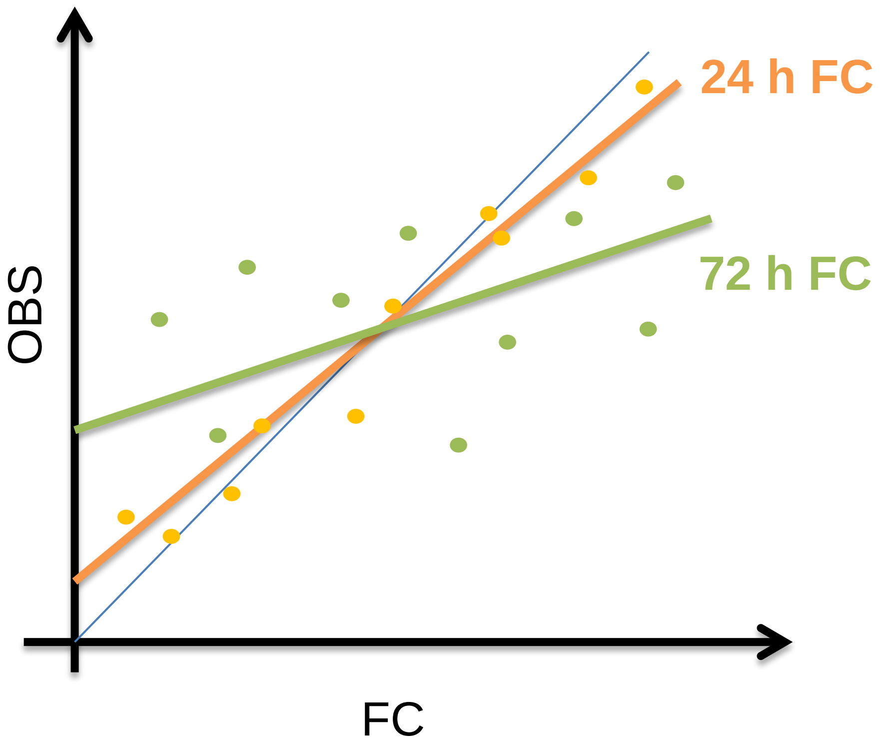
Perfect forecast = obs, points should be on the 45° diagonal

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

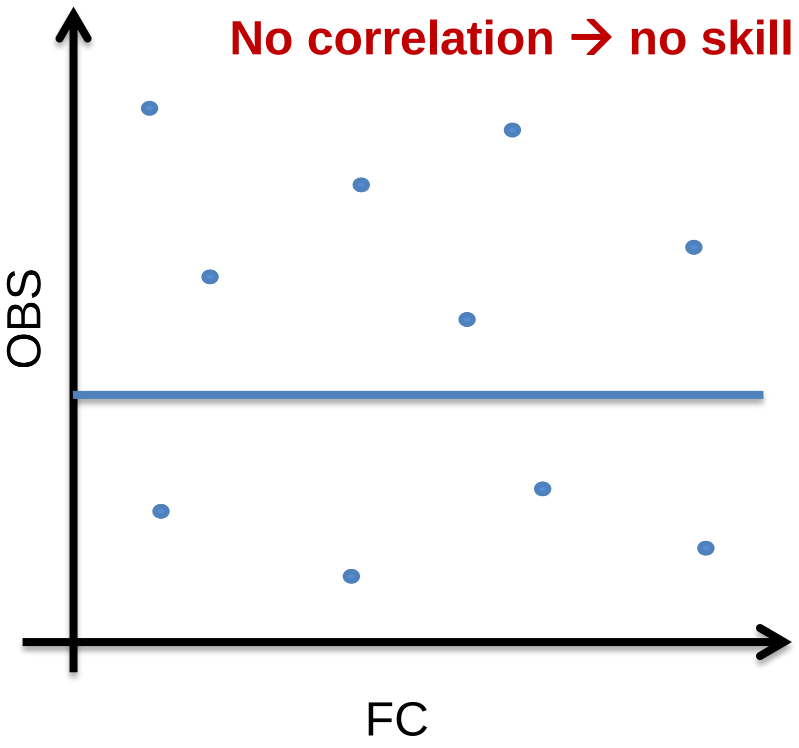


Questions:

Scatter-plot: How will the scatter plot and regression line change for longer forecasts?



Scatter-plot: How would you interpret a horizontal regression line?

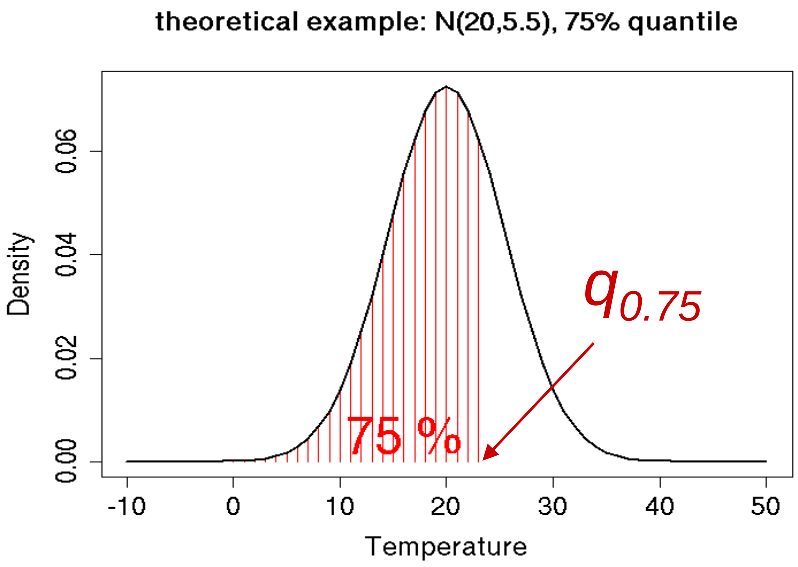


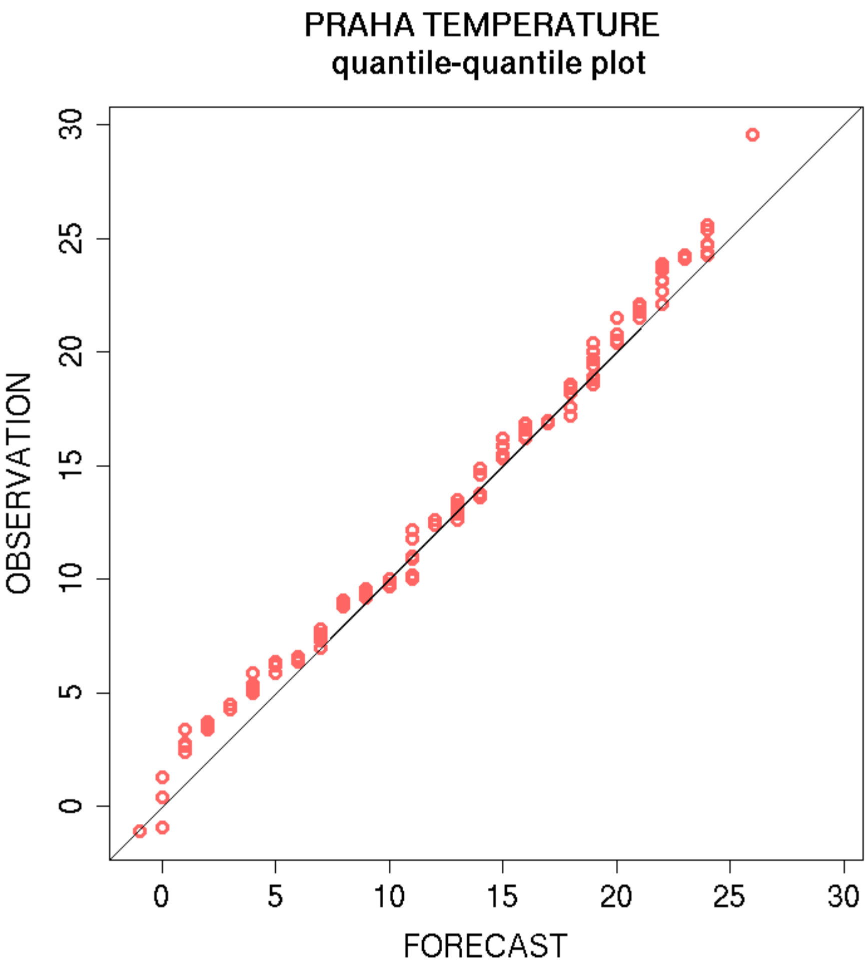
Exploratory methods: marginal distribution

Quantile-quantile plots:

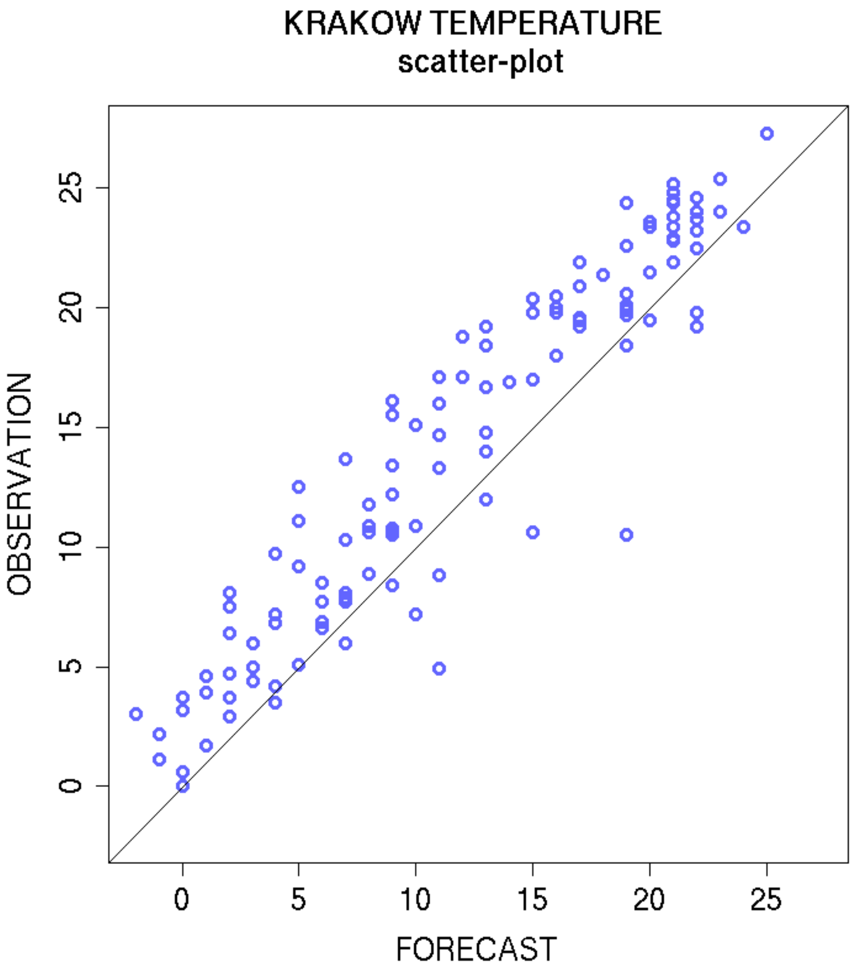
OBS quantile versus the corresponding FCST quantile

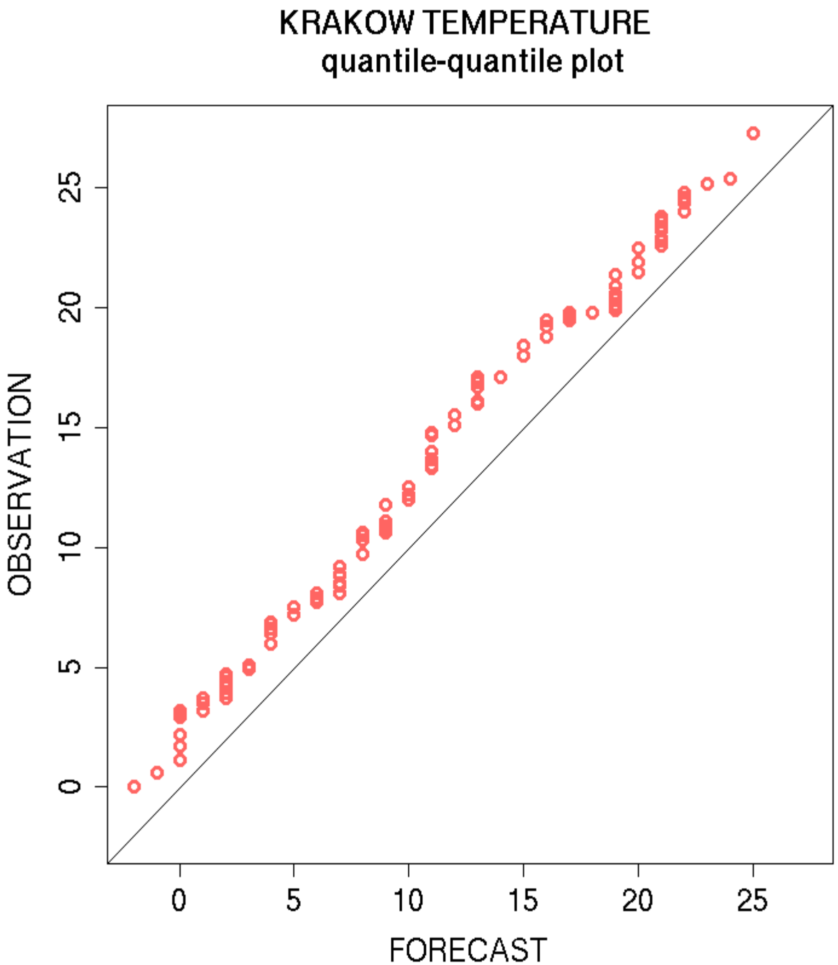
Perfect: FCST=OBS, points should be on the 45º diagonal





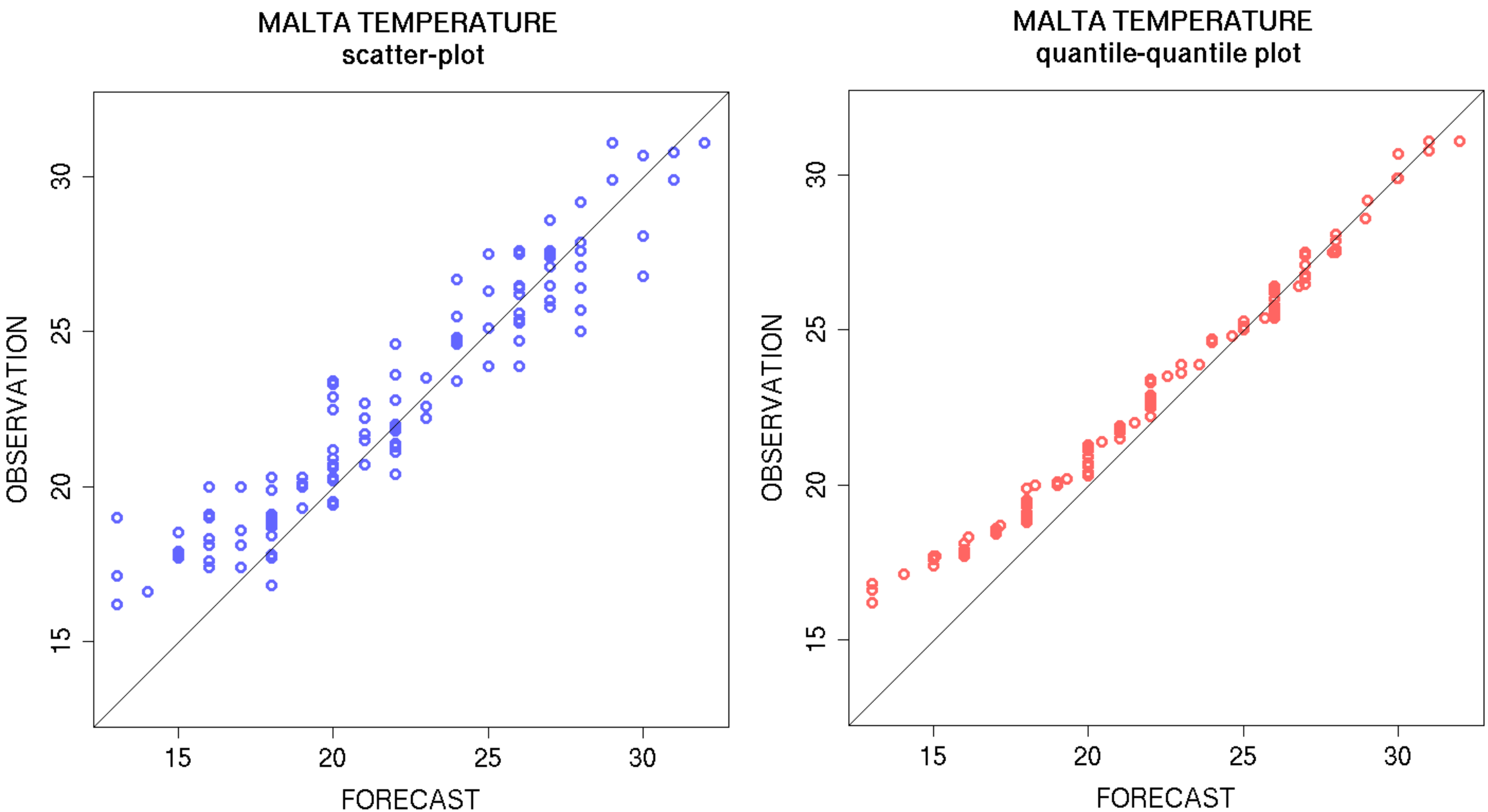
Scatter-plot and qq-plot: example 1
Q: is there any bias? Positive (over-forecast) or negative (under-forecast)?





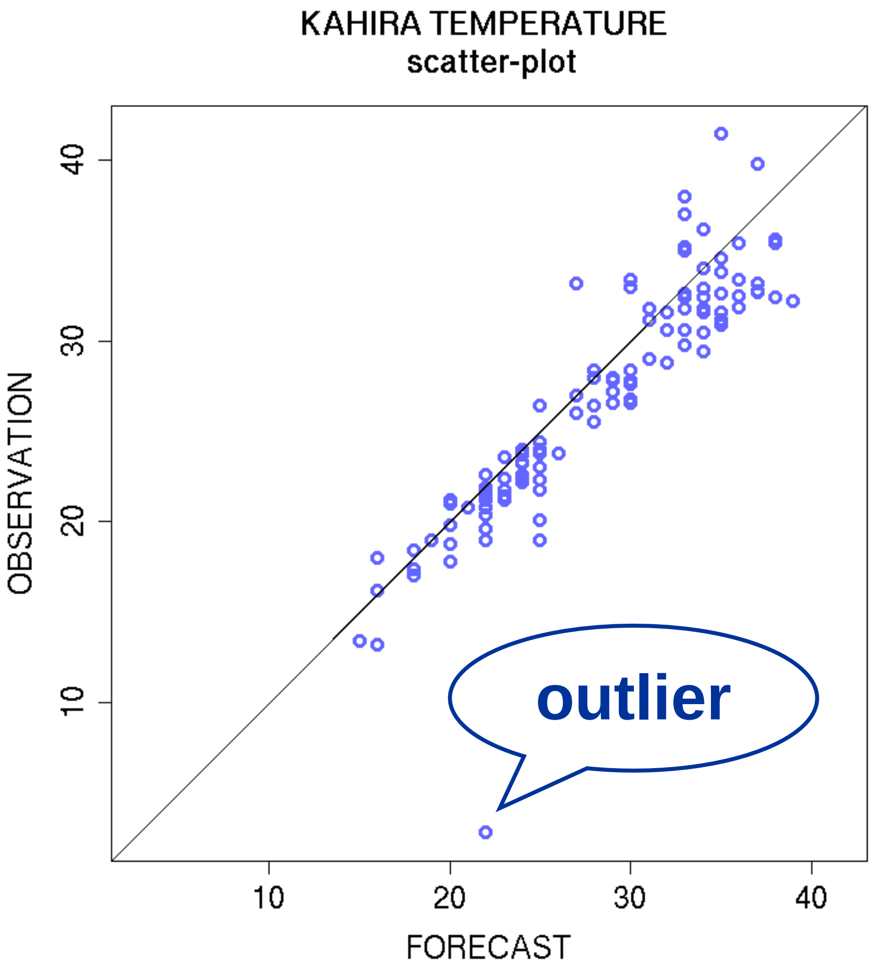
Scatter-plot and qq-plot: example 2

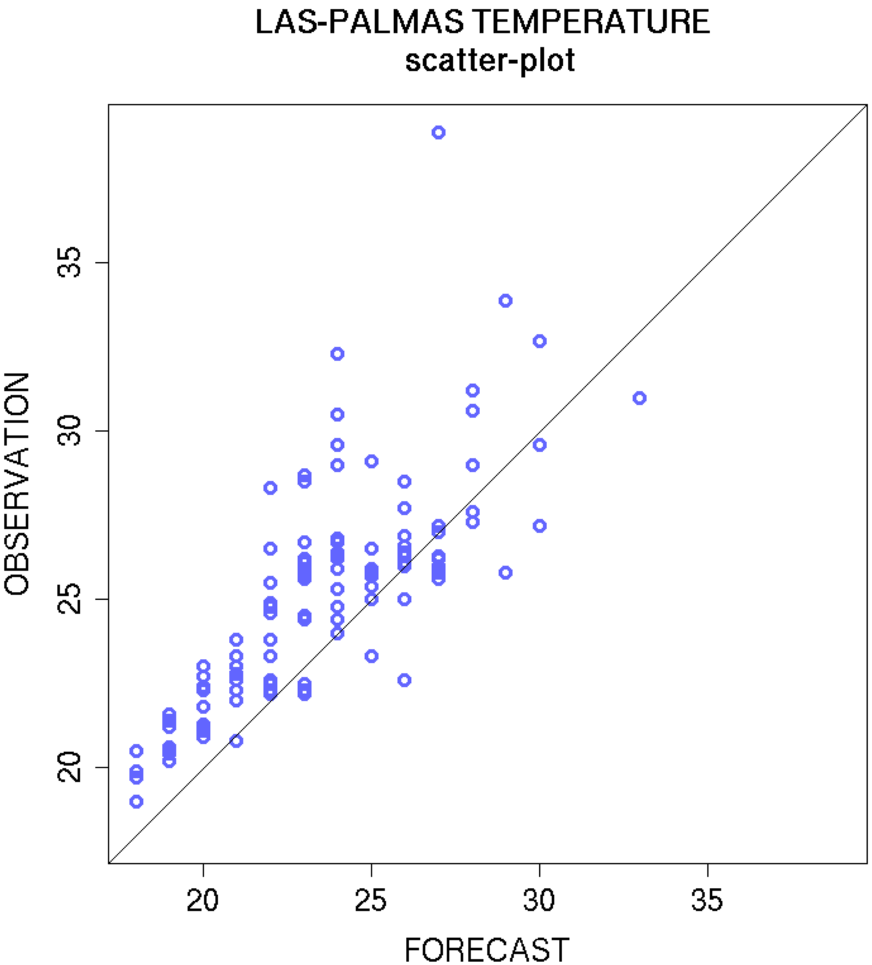
Describe the peculiar behaviour of low temperatures



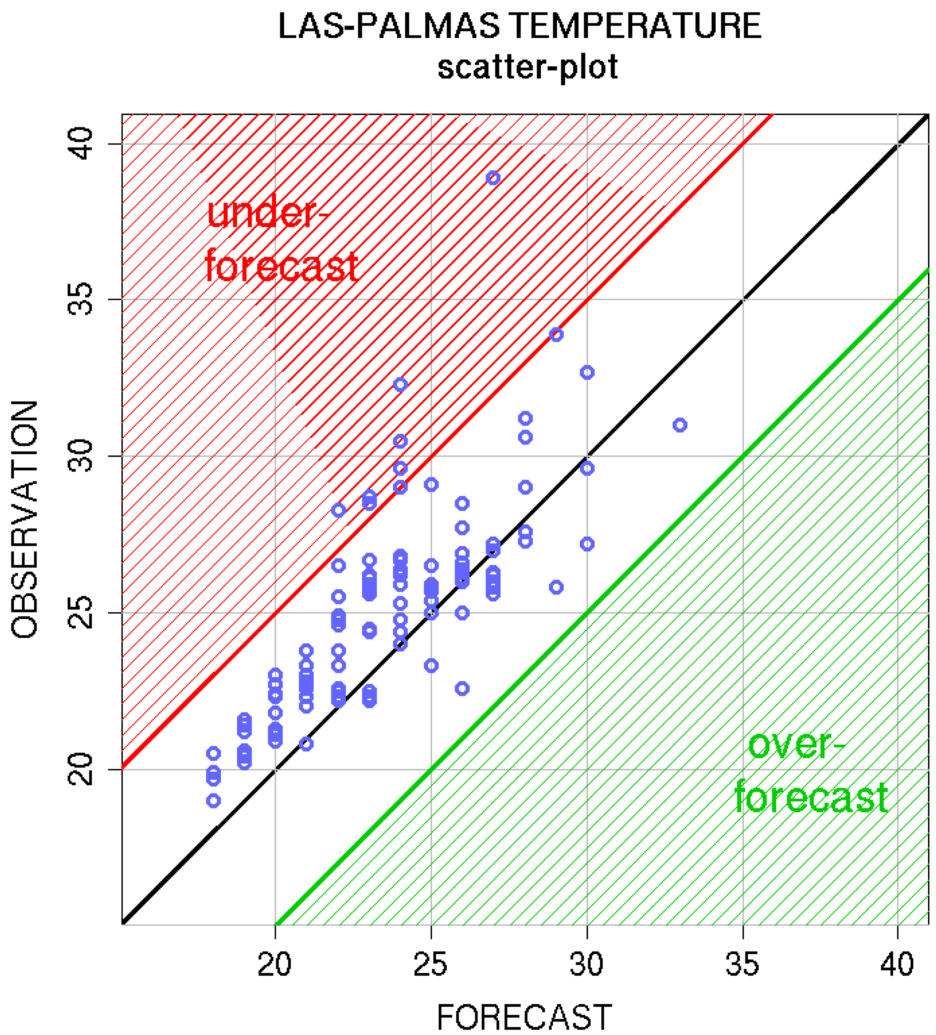
Scatter-plot: example 3

Describe how the error varies as the temperatures grow





Scatter-plot: example 4
Quantify the error



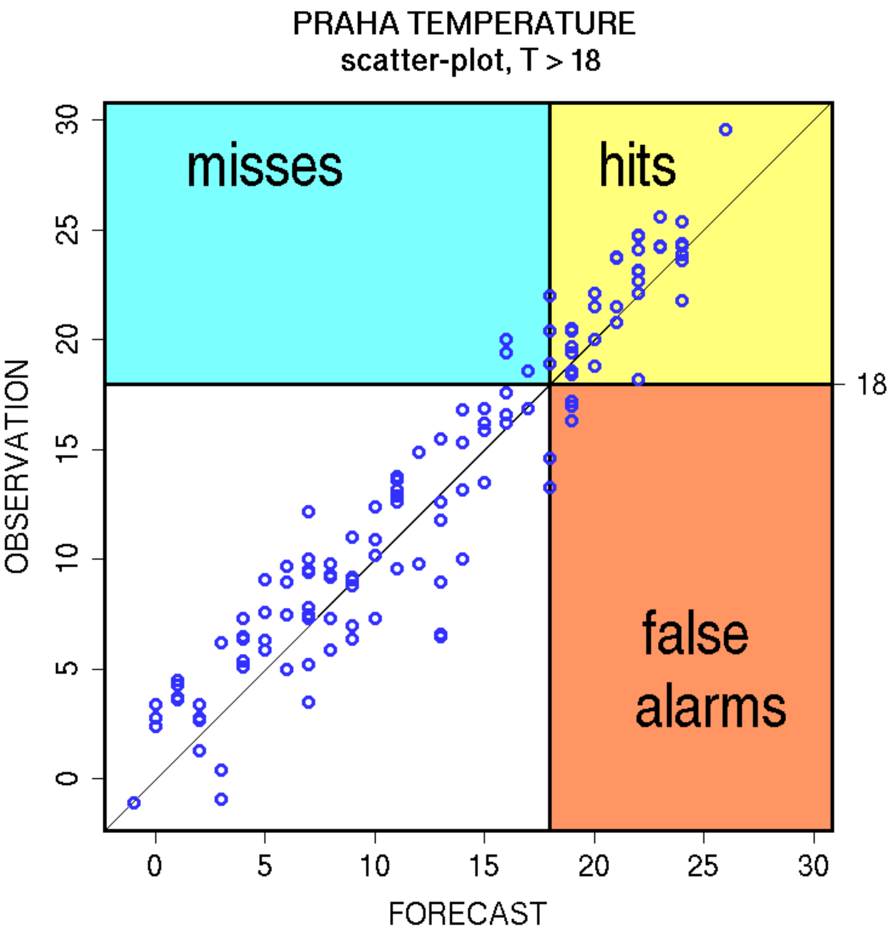
Q: how many forecasts exhibit an error larger than 10 degrees ?

Q: How many forecasts exhibit an error larger than 5 degrees ?

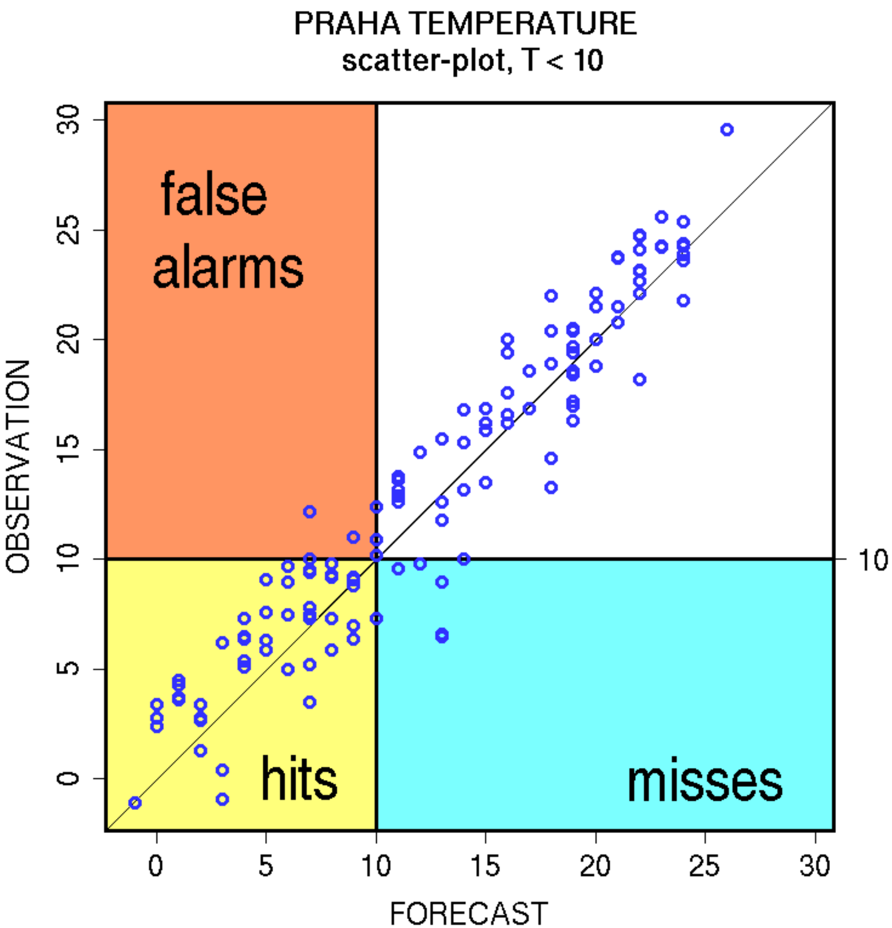
Q: Is the forecast error due mainly to an under-forecast or an over-forecast ?

Scatter-plot and Contingency Table

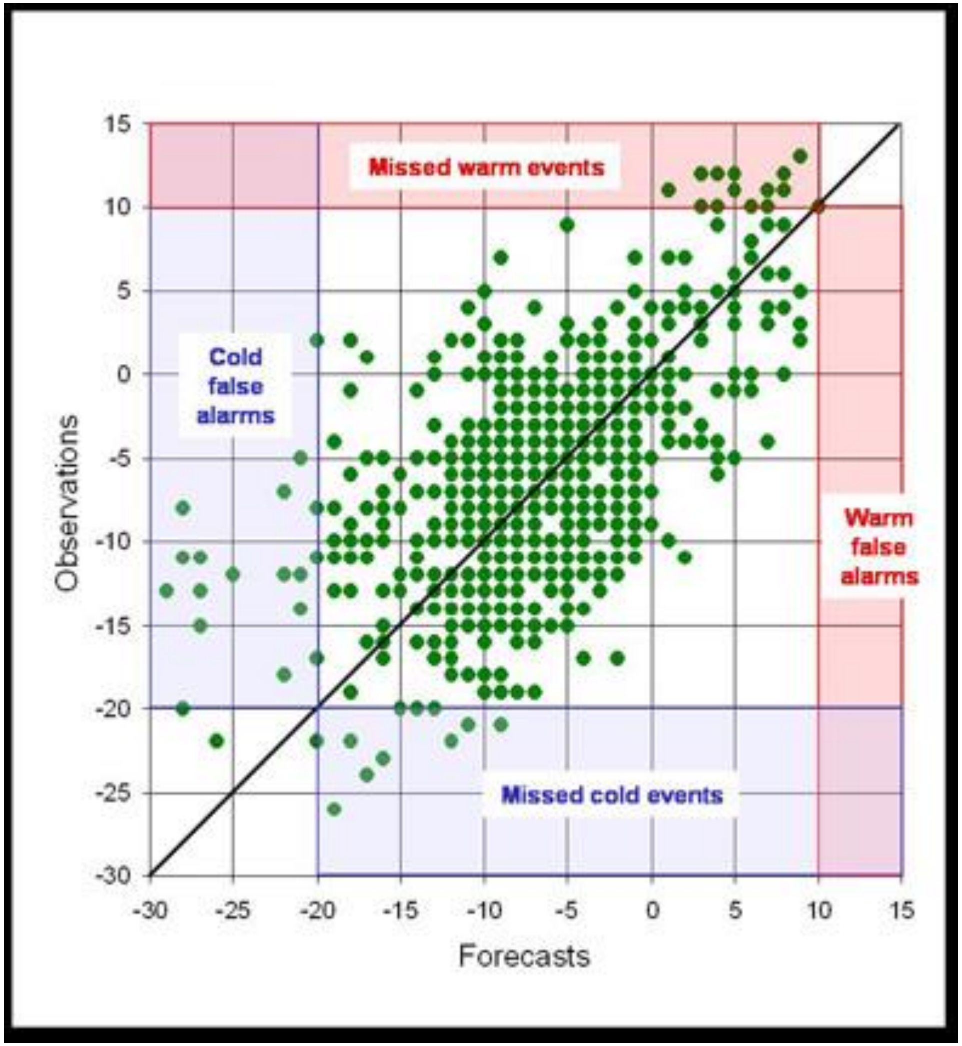
Does the forecast detect correctly temperatures above 18 degrees ?



Does the forecast detect correctly temperatures below 10 degrees ?



Scatter-plot and Cont. Table: example 5
Analysis of the extreme behavior



Q: How does the forecast handle the **temperatures above 10 degrees** ?

- How many misses ?
- How many False Alarms ?
- Is there an under- or over-forecast of temperatures larger than 10 degrees ?

Q: How does the forecast handle the **temperatures below -20 degrees** ?

- How many misses ?
- Are there more missed cold events or false alarms cold events ?
- How does the forecast minimum temperature compare with the observed minimum temperature ?

Exploratory methods: marginal distributions

Visual comparison:
Histograms, box-plots, …

Summary statistics:

- ***Location:***

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

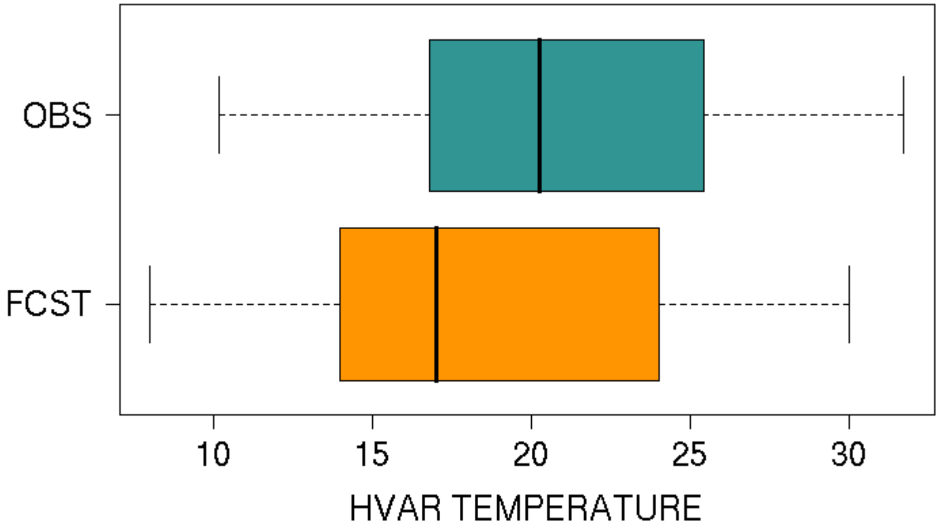
$$\text{median} = q_{0.5}$$

- ***Spread:***

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

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

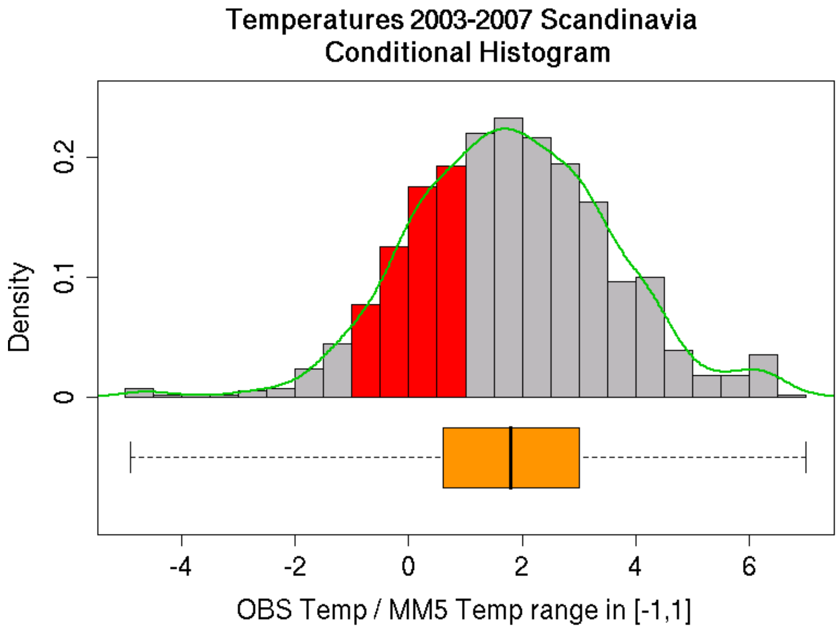
$$\text{IQR} = q_{0.75} - q_{0.25}$$

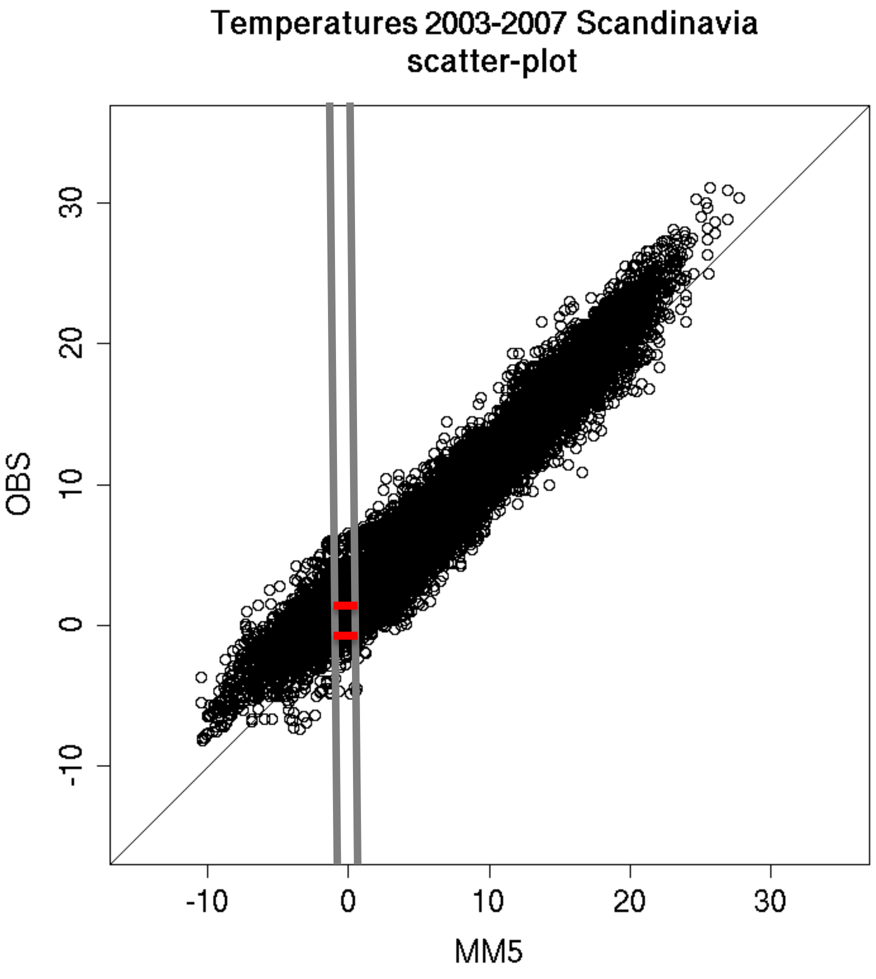


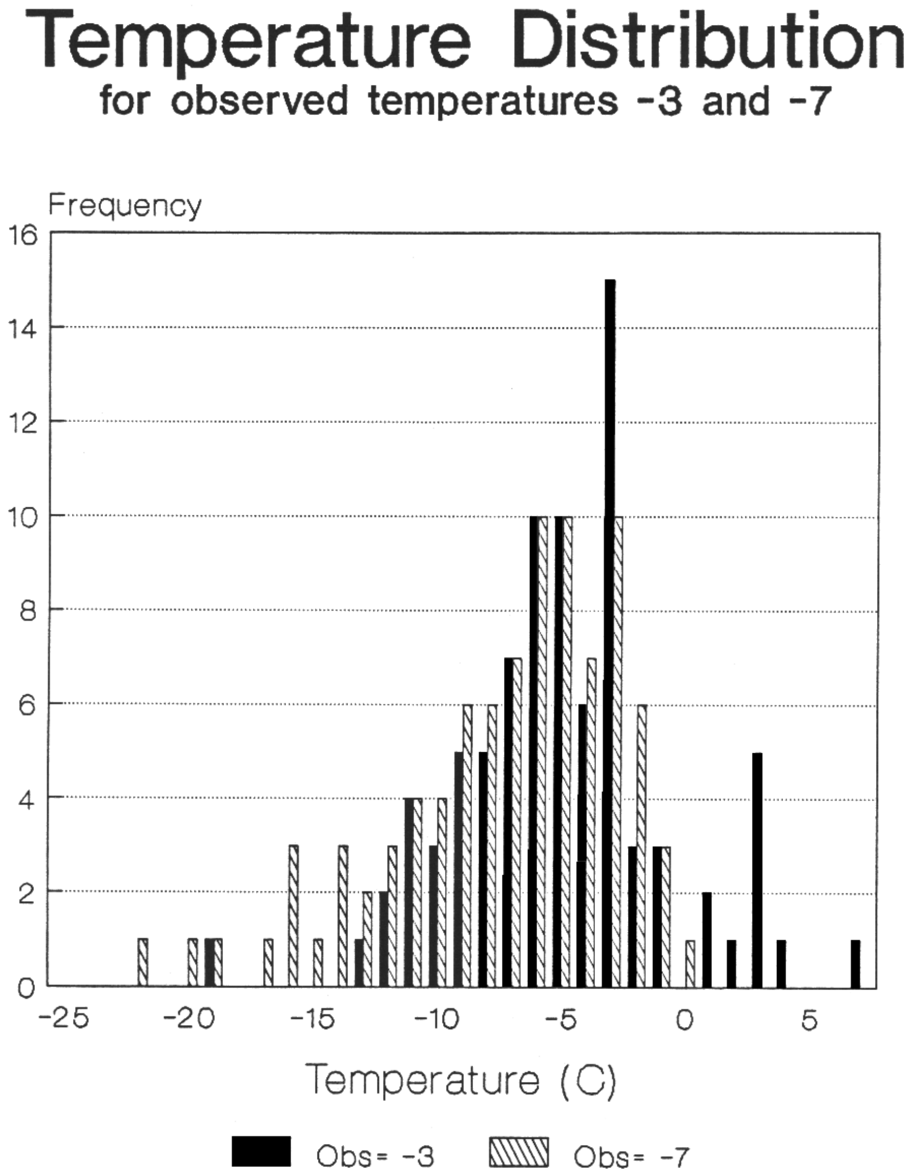
	MEAN	MEDIAN	STDEV	IQR
OBS	20.71	20.25	5.18	8.52
FCST	18.62	17.00	5.99	9.75

Exploratory methods: conditional distributions

Conditional histogram and conditional box-plot







Q: Look at the figure: What can you say about the forecast system?

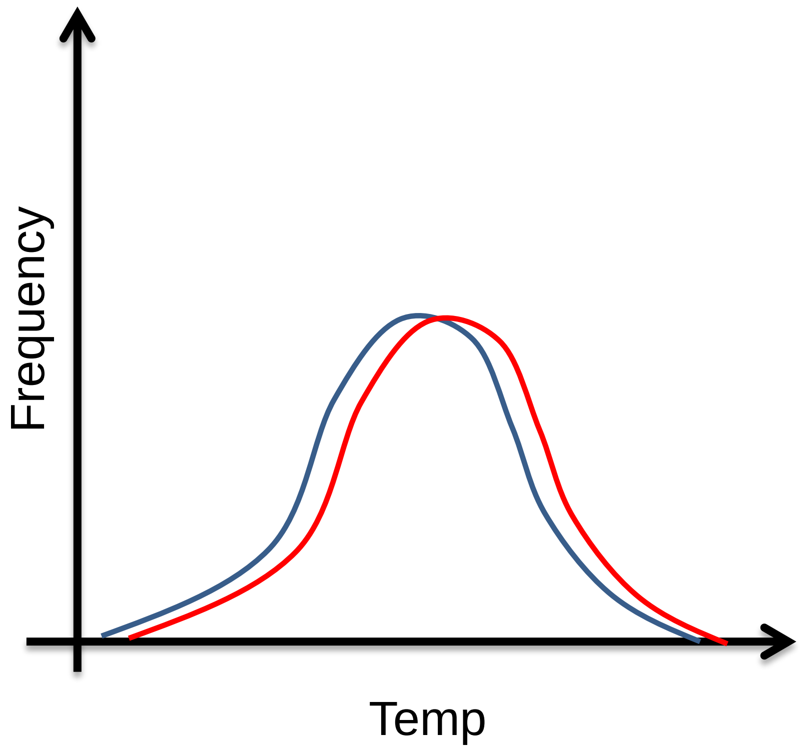
→ cannot discriminate

Histogram of forecast temperatures given an observed temperature of -3 deg C and -7 deg C. 11 Atlantic region stations for the period 1/86 to 3/86. Sample size 701 cases.

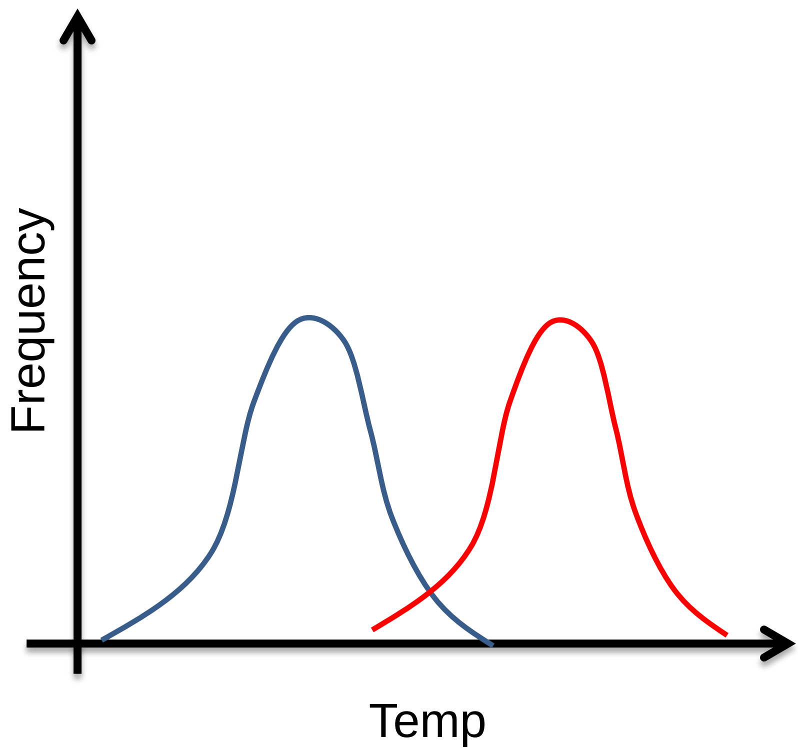
Stanski et al., 1989

Exploratory methods: conditional distributions

cannot discriminate



can discriminate



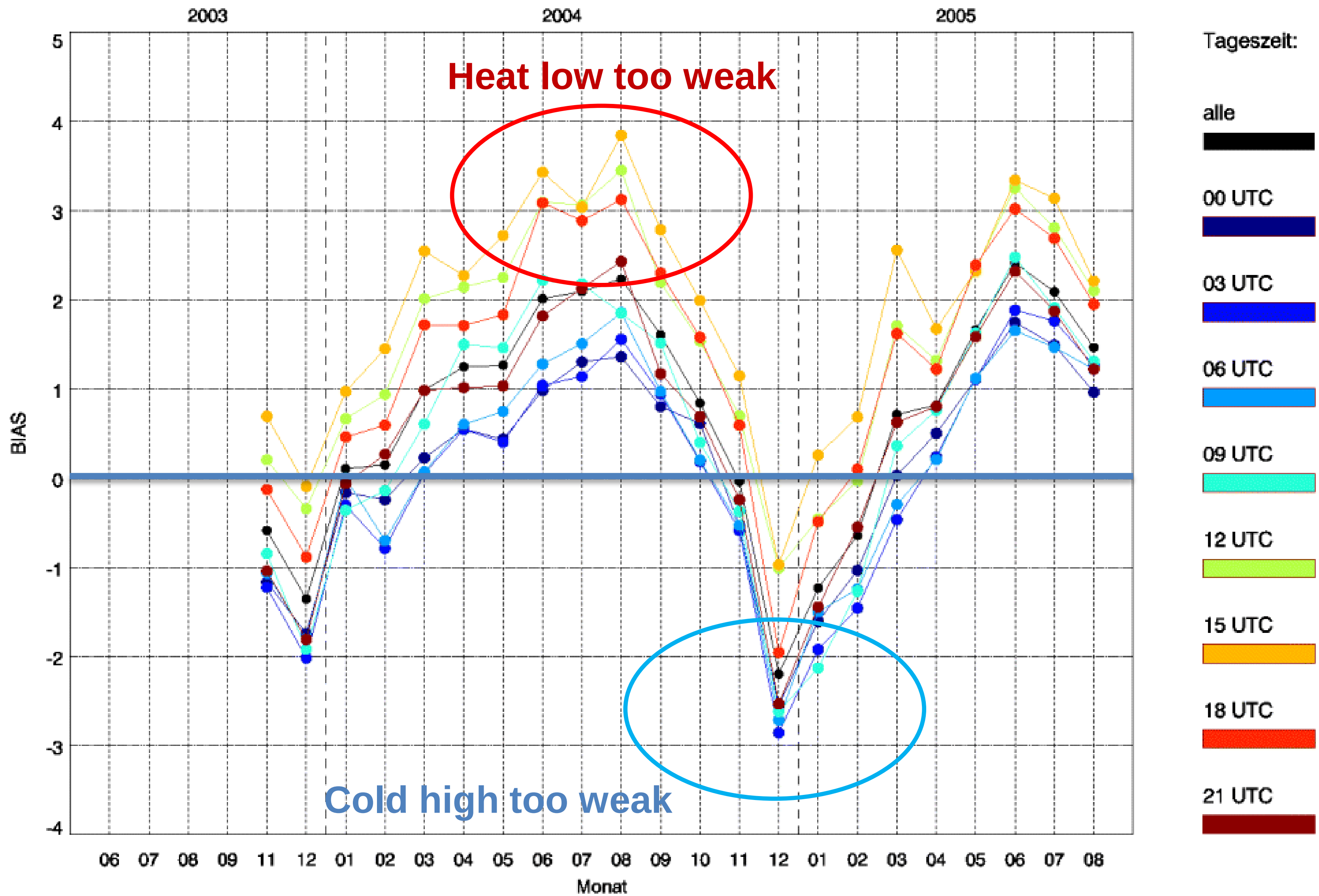
Scores for continuous forecasts: linear bias

$$\text{Bias} = \text{Mean Error} = ME = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - x_i\right) = \overline{f} - \overline{x}$$

f = forecast; x = observation

- Measures the **average of the errors** = difference between the forecast and observed means
- Indicates the average *direction* of error: positive bias indicates over-forecast, negative bias indicates under-forecast (→ bias correction)
- Does *not* indicate the *magnitude* of the error (positive and negative error can – and hopefully do – cancel out)

Monthly mean bias of MSLP field (LM-VERA) in hPa over eastern Alps



Gorgas, 2006

Scores for continuous forecasts: Mean Absolute Error (MAE)

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

- 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

Continuous scores: MSE

$$\text{Mean Squared Error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - x_i\right)^2$$

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

Continuous scores: RMSE

$$\text{RMSE} = \sqrt{\text{MSE}}$$

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

Q: if I verify two sets of data and in one I find RMSE $\gg$ MAE, in the other I find RMSE $\gtrsim$ MAE, which set is more likely to have large outliers ? Which set has larger variance ?

Continuous scores: linear correlation

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

Attribute: measures association

Measures linear association between forecast and observation

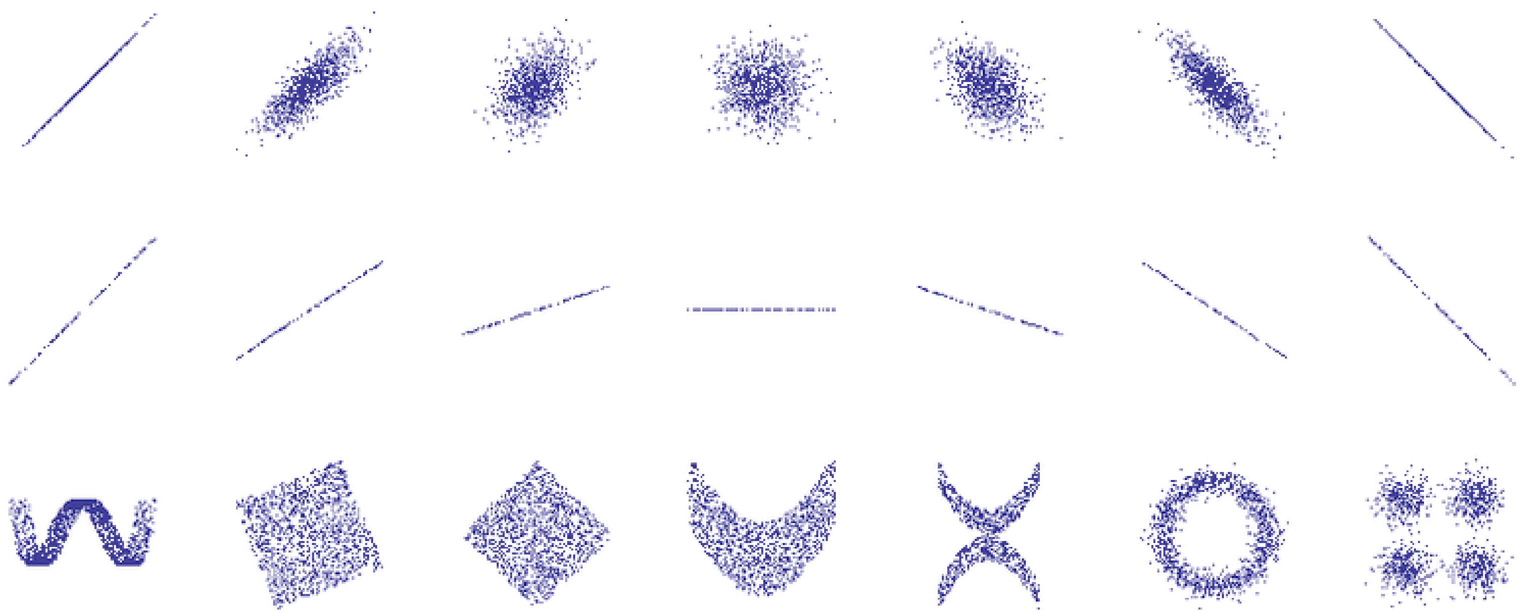
Y and X rescaled (non-dimensional) covariance: ranges in [-1,1]

It is not sensitive to the bias

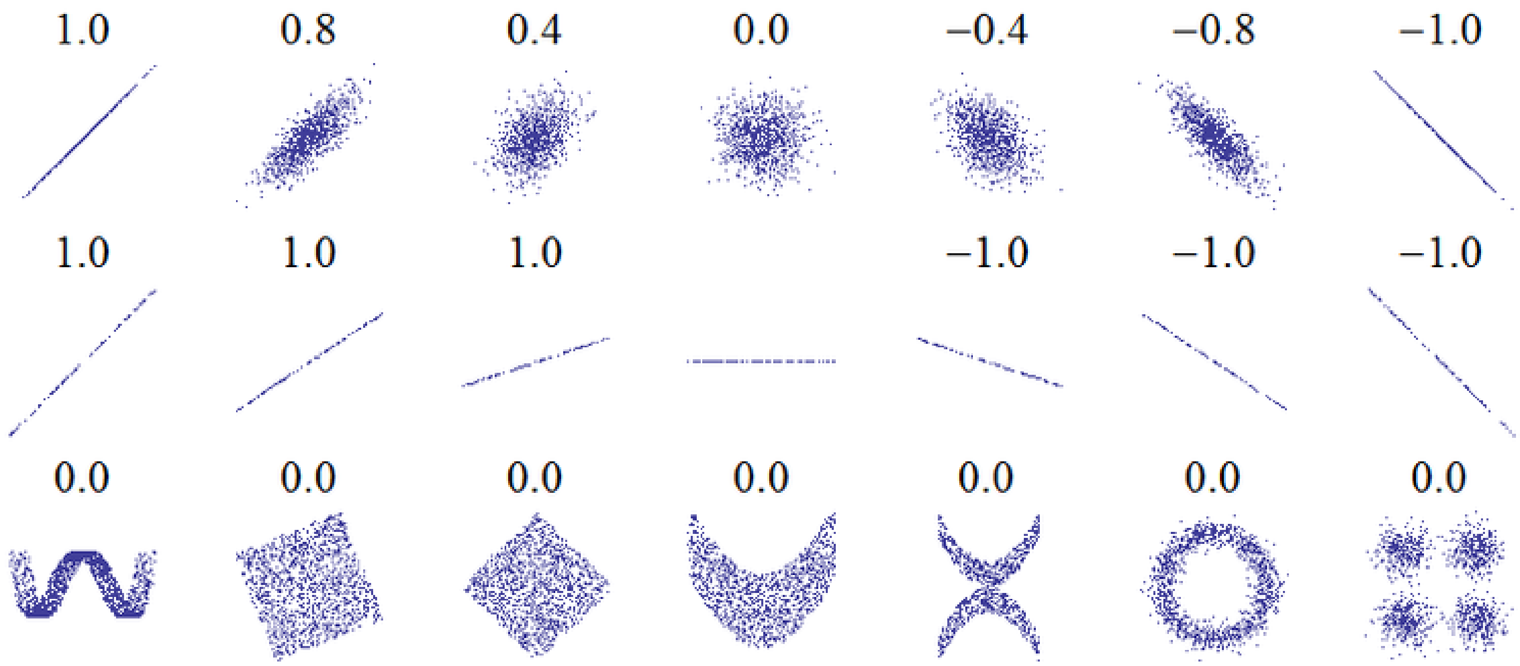
The correlation coefficient alone does not provide information on the inclination of the regression line (it says only is it is positively or negatively tilted); observation and forecast variances are needed; the slope coefficient of the regression line is given by $b = (s_X/s_Y)r_{XY}$

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

Correlation coefficient

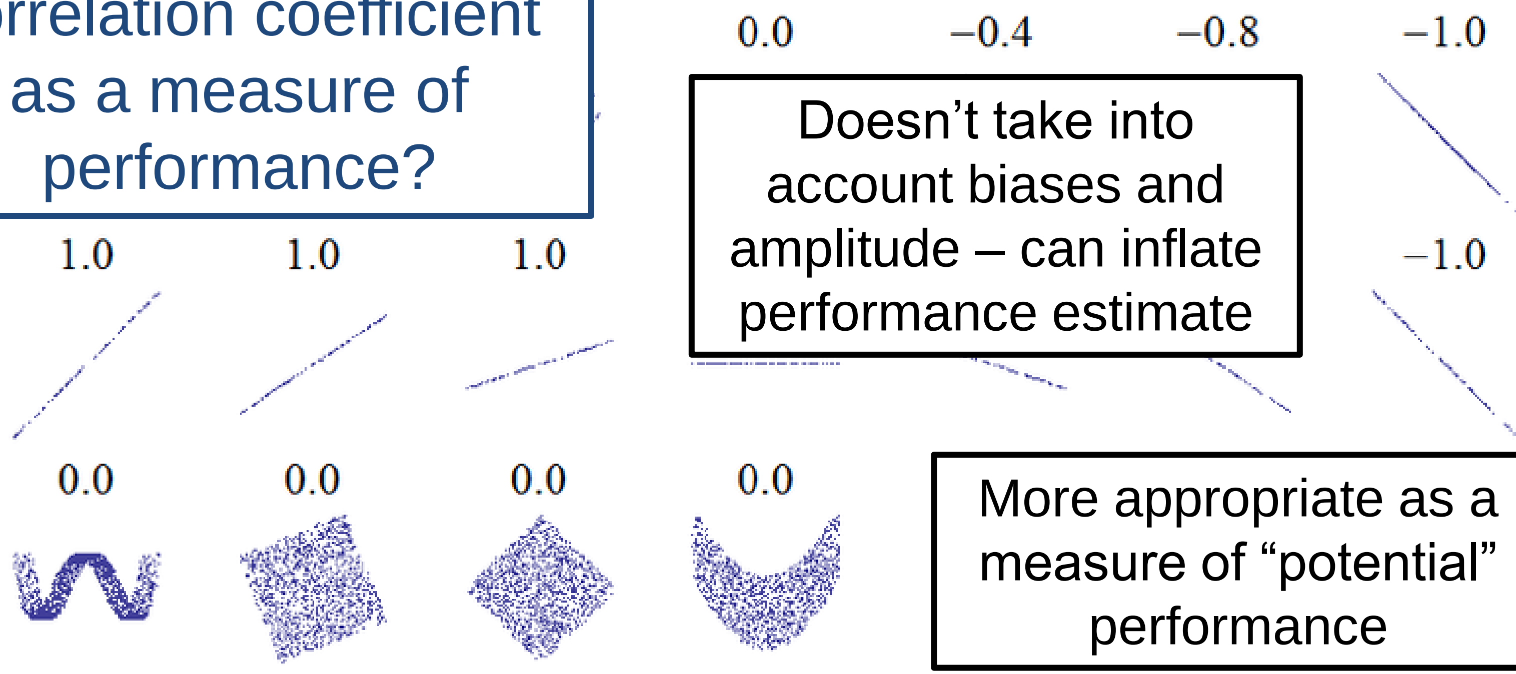


Correlation coefficient



Correlation coefficient

What is wrong with the correlation coefficient as a measure of performance?



Doesn’t take into account biases and amplitude – can inflate performance estimate

More appropriate as a measure of “potential” performance

Decomposition of the MSE

$$f = f' + \overline{f}$$

$$o = o' + \overline{o}$$

→ **Reynold's Averaging**

$$\overline{f'} = 0$$

$$\overline{o'} = 0$$

$$MSE = \overline{(f - o)^2}$$

$$MSE = \overline{f'^2} + \overline{o'^2} + \left(\overline{f} - \overline{o}\right)^2 - 2\overline{f'o'}$$

$$MSE = \sigma_f^2 + \sigma_o^2 + bias^2 - 2 * \mathrm{cov}(f, o)$$

$$\boxed{MSE = \sigma_f^2 + \sigma_o^2 + bias^2 - 2 * \sigma_f \sigma_o cor(f, o)}$$

Bias can be subtracted !
BC_(R)MSE

Consequence: smooth forecasts verify better

$$MSE \overset{!}{=} \min$$

$$\frac{\partial MSE}{\partial \sigma_f} \overset{!}{=} 0$$

$$\boxed{\sigma_{f\_MSE\_optimal} = \sigma_o cor(f, o)}$$

Taylor Diagramm

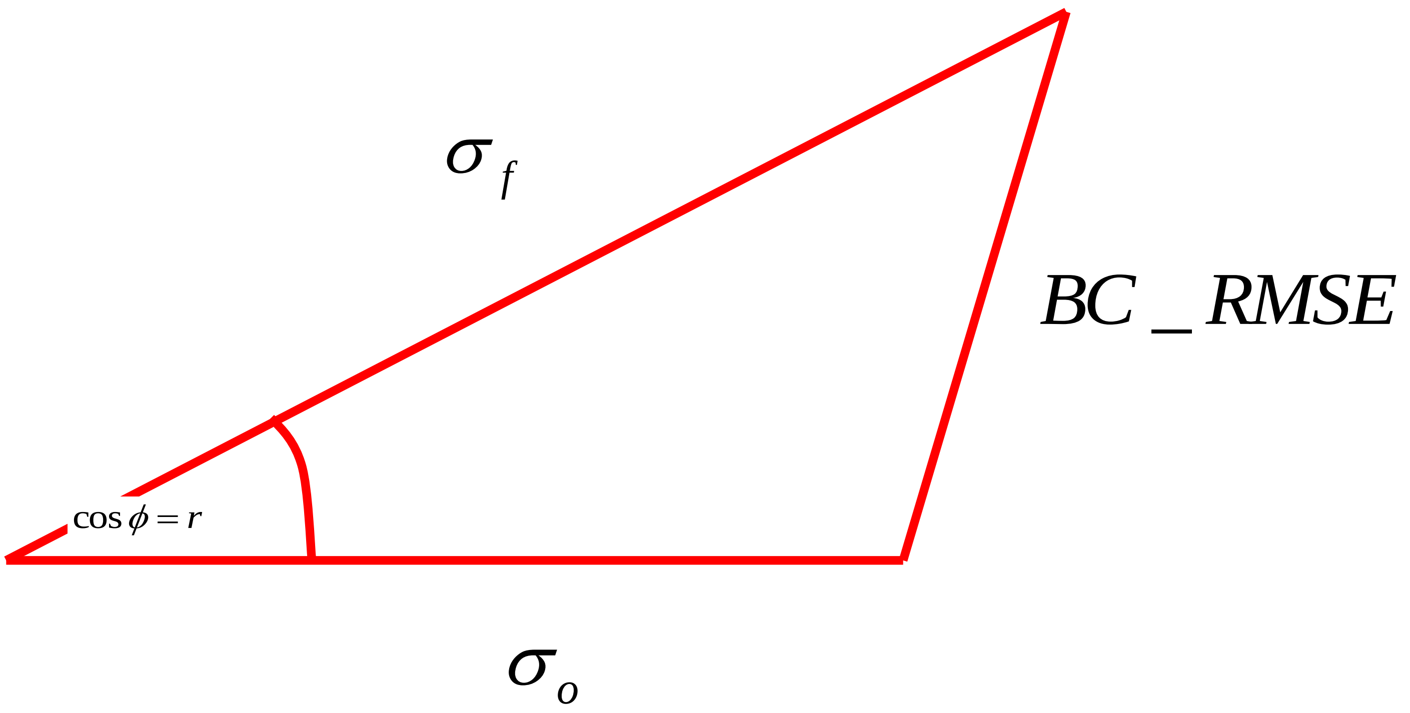
Combines BC_RMSE, variance and correlation coefficient in a graphical way

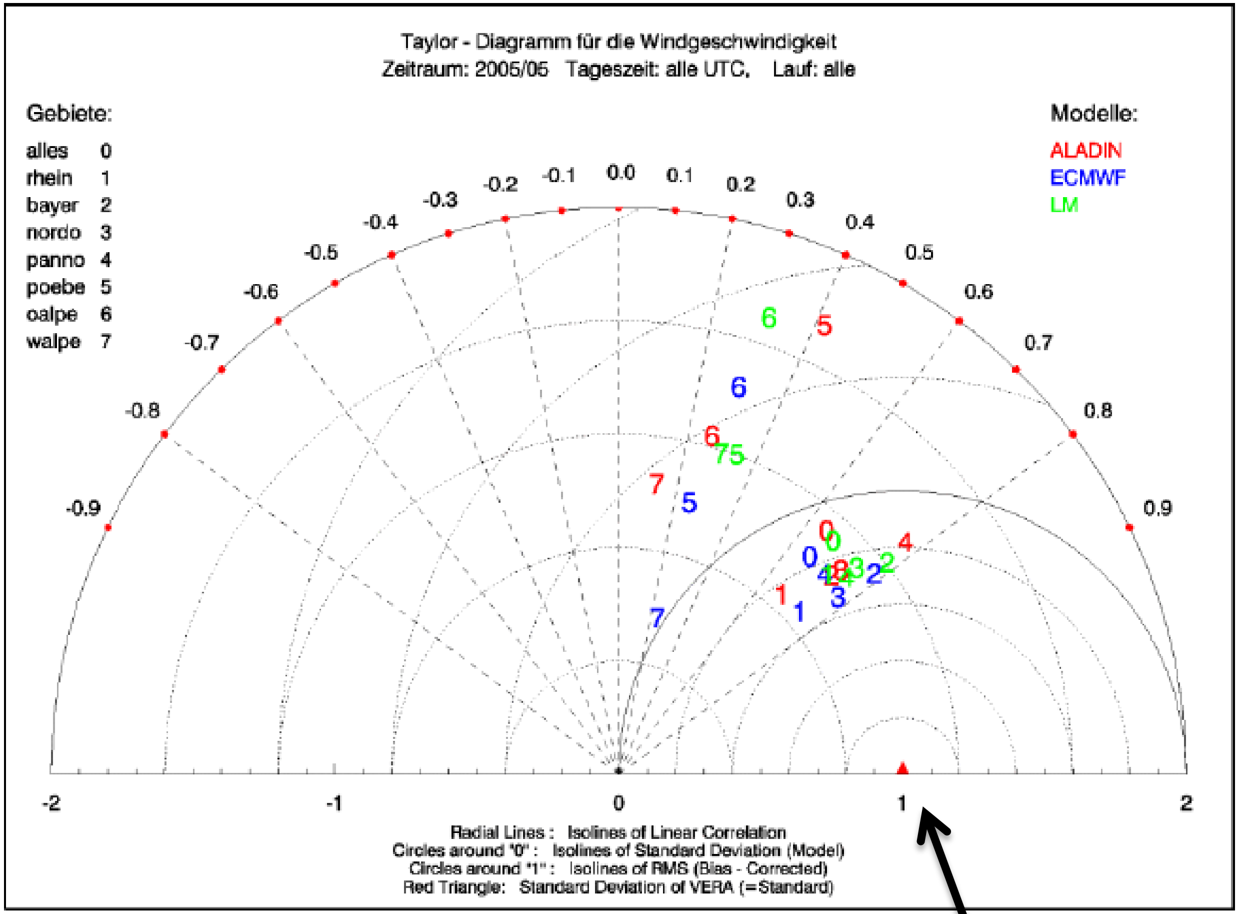
$$BC\_RMSE^2 = \frac{1}{N}\sum \left[\left(X^f - \overline{X}^f\right) - \left(X^o - \overline{X}^o\right)\right]^2$$

$$BC\_RMSE^2 = \sigma_f^2 + \sigma_o^2 - 2\sigma_f \sigma_o r$$

$$r = \frac{\mathrm{cov}(X^f, X^o)}{\sigma^f \sigma^o}$$

Law of cosines: $$c^2 = a^2 + b^2 - 2a\,b\,\cos\phi$$


σ_f
BC_RMSE
$\cos\phi = r$
σ_o



Gorgas, 2006

Reference

Comparative verification

Skill scores

- A skill score is a measure of *relative performance*
  - ***Ex***: *How much more accurate are my temperature predictions than climatology? How much more accurate are they than the model's temperature predictions?*
  - *Provides a comparison to a* ***standard***
- *Standard of comparison (=reference) can be*
  - *Chance (easy?)*
  - *Long-term climatology (more difficult)*
  - *Sample climatology (difficult)*
  - *Competitor model / forecast (most difficult)*
  - *Persistence (hard or easy)*

Comparative verification

– Generic skill score definition:

$$SS = \frac{M - M_{ref}}{M_{perf} - M_{ref}}$$

Where $\boldsymbol{M}$ is the verification measure for the forecasts, $\boldsymbol{M_{ref}}$ is the measure for the reference forecasts, and $\boldsymbol{M_{perf}}$ is the measure for perfect forecasts (=0)

– Measures percent improvement of the forecast over the reference

– Positively oriented (larger is better)

– Choice of the standard matters (*a lot*!) → have in mind when comparing skill scores

– Perfect score: 1

– How far I am on the way to the perfect forecast?

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:

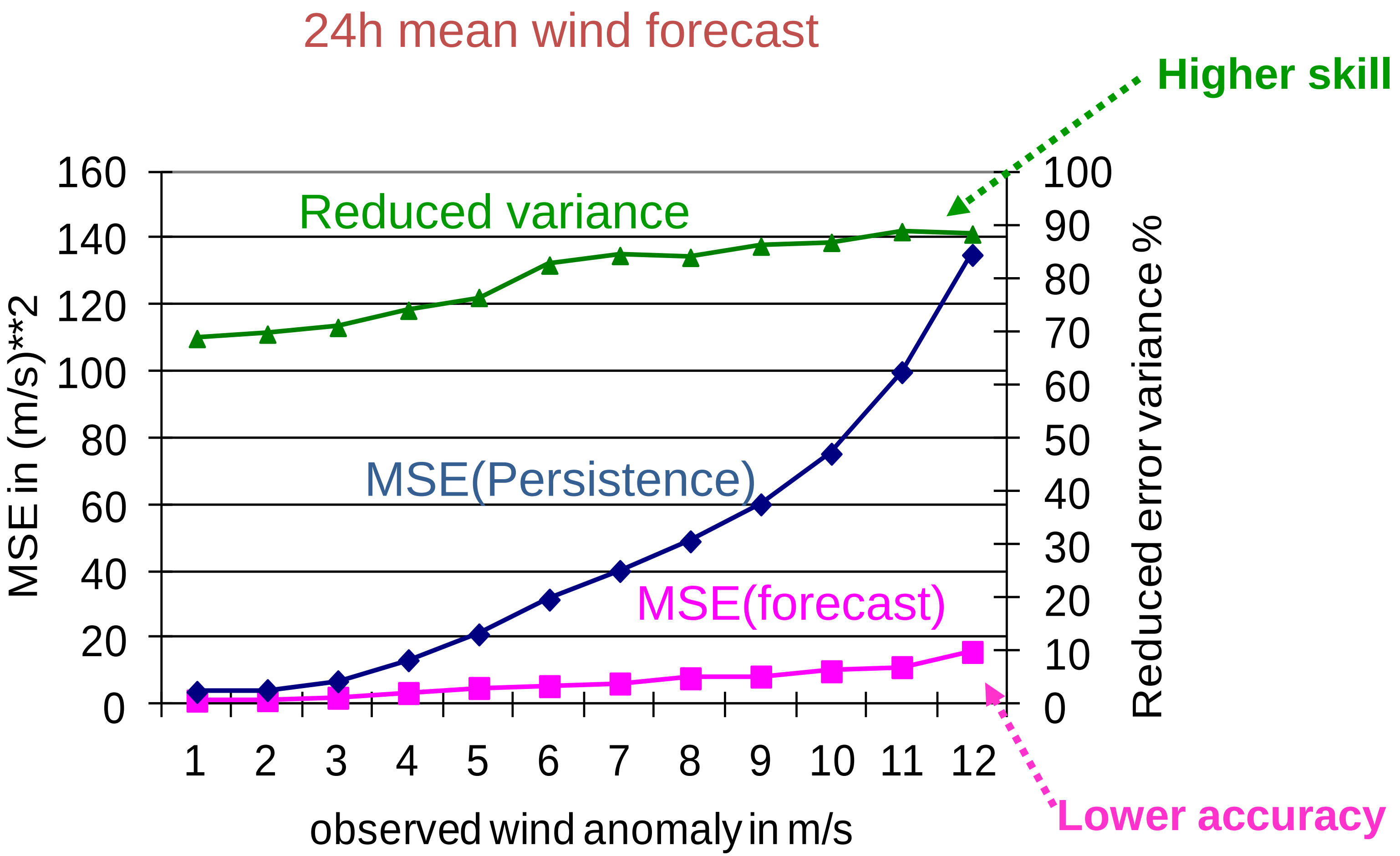
linear correlation

bias

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

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

Accuracy vs skill



→ **High skill because getting reference worse.**

Continuous scores: anomaly correlation

$$y'_m = y_m - c_m$$
$$x'_m = x_m - c_m$$

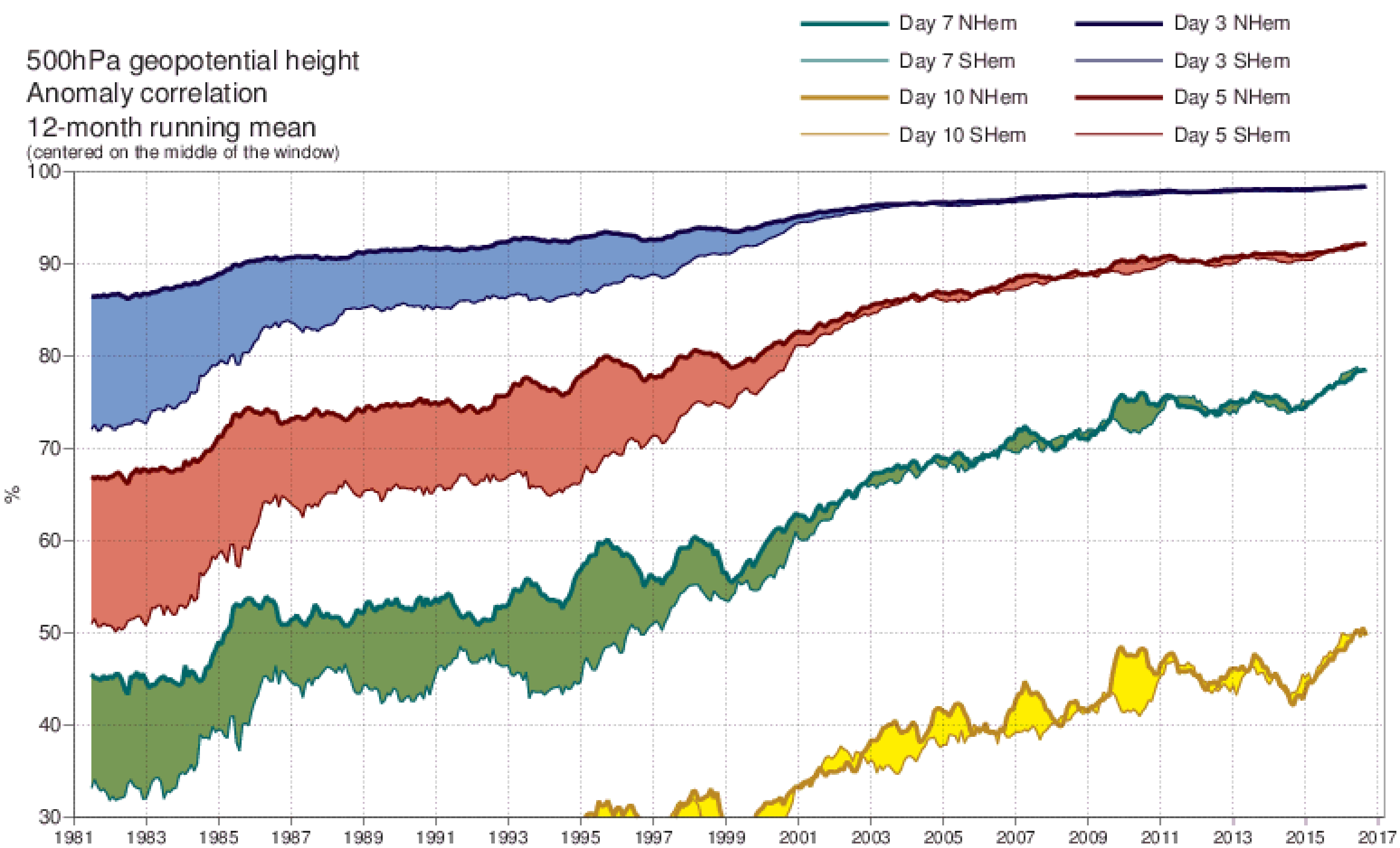
Forecast and observation anomalies to evaluate forecast quality not accounting for correct forecast of climatology (e.g. driven by topography)

$$AC_{cent} = \frac{\sum_{m \in map} \left(y'_m - \overline{y}' \right)\left(x'_m - \overline{x}' \right)}{\sqrt{\sum_{m \in map} \left(y'_m - \overline{y}' \right)^2 \sum_{m \in map} \left(x'_m - \overline{x}' \right)^2}}$$

Centred and uncentred AC **for weather variables defined over a spatial domain**: c_m is the climatology at the grid-point m, over-bar denotes averaging over the field

$$AC_{unc} = \frac{\sum_{m \in map} \left(y_m - c_m \right)\left(x_m - c_m \right)}{\sqrt{\sum_{m \in map} \left(y_m - c_m \right)^2 \sum_{m \in map} \left(x_m - c_m \right)^2}} = \frac{\sum_{m \in map} \left(y'_m \right)\left(x'_m \right)}{\sqrt{\sum_{m \in map} \left(y'_m \right)^2 \sum_{m \in map} \left(x'_m \right)^2}}$$

Continuous scores: anomaly correlation

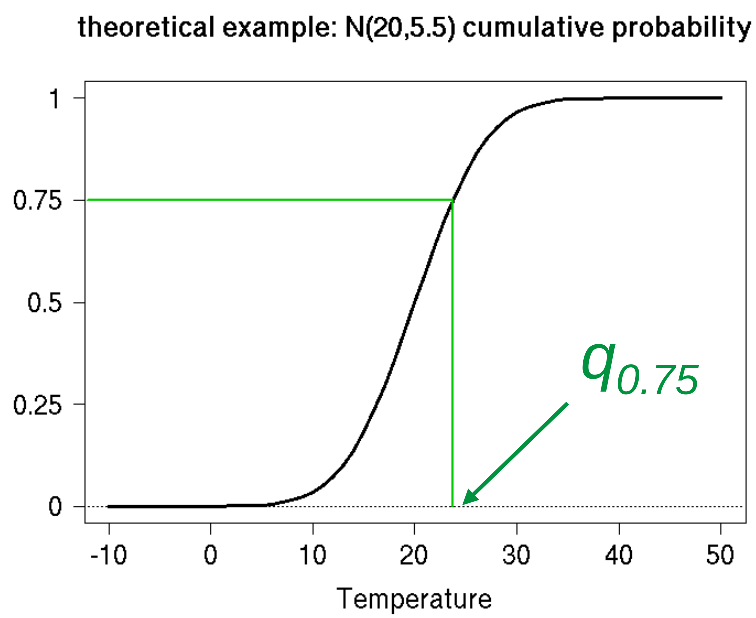


ECMWF

Linear Error in Probability Space

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

- LEPS is an 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
- More robust (equitable) version developed by Potts (1996)



Summary

- Graphical representations of distributions provide a great deal of information about performance
  - Use initially to characterize forecasts and observations
  - Can also be used to depict performance and comparative performance
- Joint, marginal, and conditional distributions provide different kinds of information
  - Summary scores and measures also provide different kinds of information

Summary cont.

- Many summary scores exist for each type of distribution
  - Each provides different kinds of information
- High dimensionality of the continuous forecast verification problem requires use of a variety of measures
- Selection of a particular standard of comparison will have a big impact on skill
  - Easy standard of comparison => Highest skill
  - Difficult standard of comparison => Lowest skill
  - Best to choose a <u>*meaningful*</u> standard

Summary cont.

- From a practical perspective:
  - *Correlation* provides limited information on its own
  - *RMSE* and *bias* are not independent
    - More meaningful to present bias-corrected RMSE along with Bias
- When planning verification give careful consideration to
  - *Sampling* (independent samples; meaningful subsets)
  - *Statistical characteristics* of forecasts and obs
  - *Performance attributes* to measure to answer questions of interest

Thank you!

References:

Jolliffe and Stephenson (2012): Forecast Verification: a practitioner's guide, 2nd Ed. Wiley & Sons.

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

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

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html