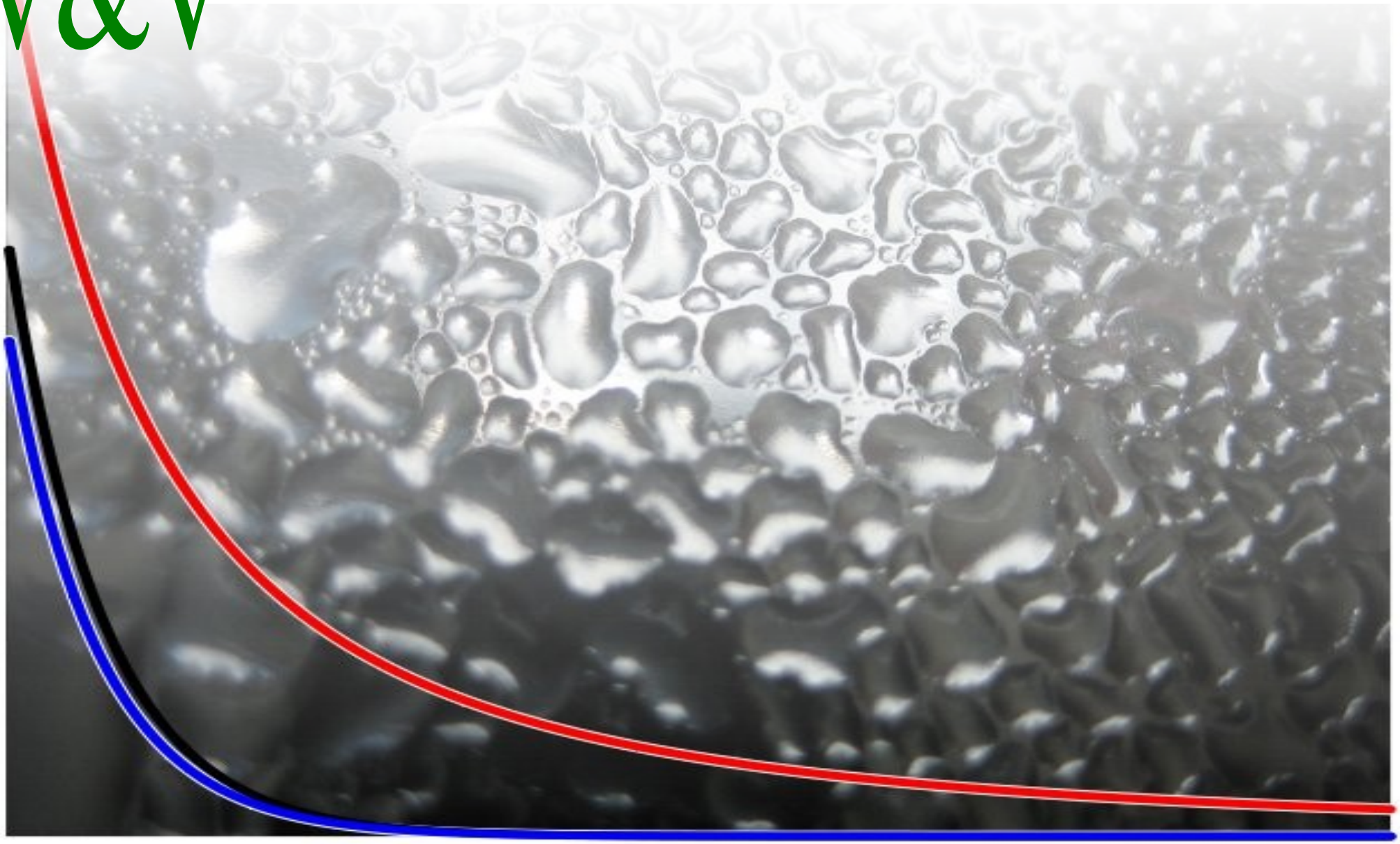
2b. Verification: statistics



After/during replication

The analysis is repeated, but...

Are the results similar?

What do the numbers reveal?
Is there a pattern?
Is it accidental?

...and what is the chance for that?

Model skill

What is our objective?

Depends on the purpose of the model. Weather model – can it tell us when/where it's going to rain? What are our expectations?

What is skill?

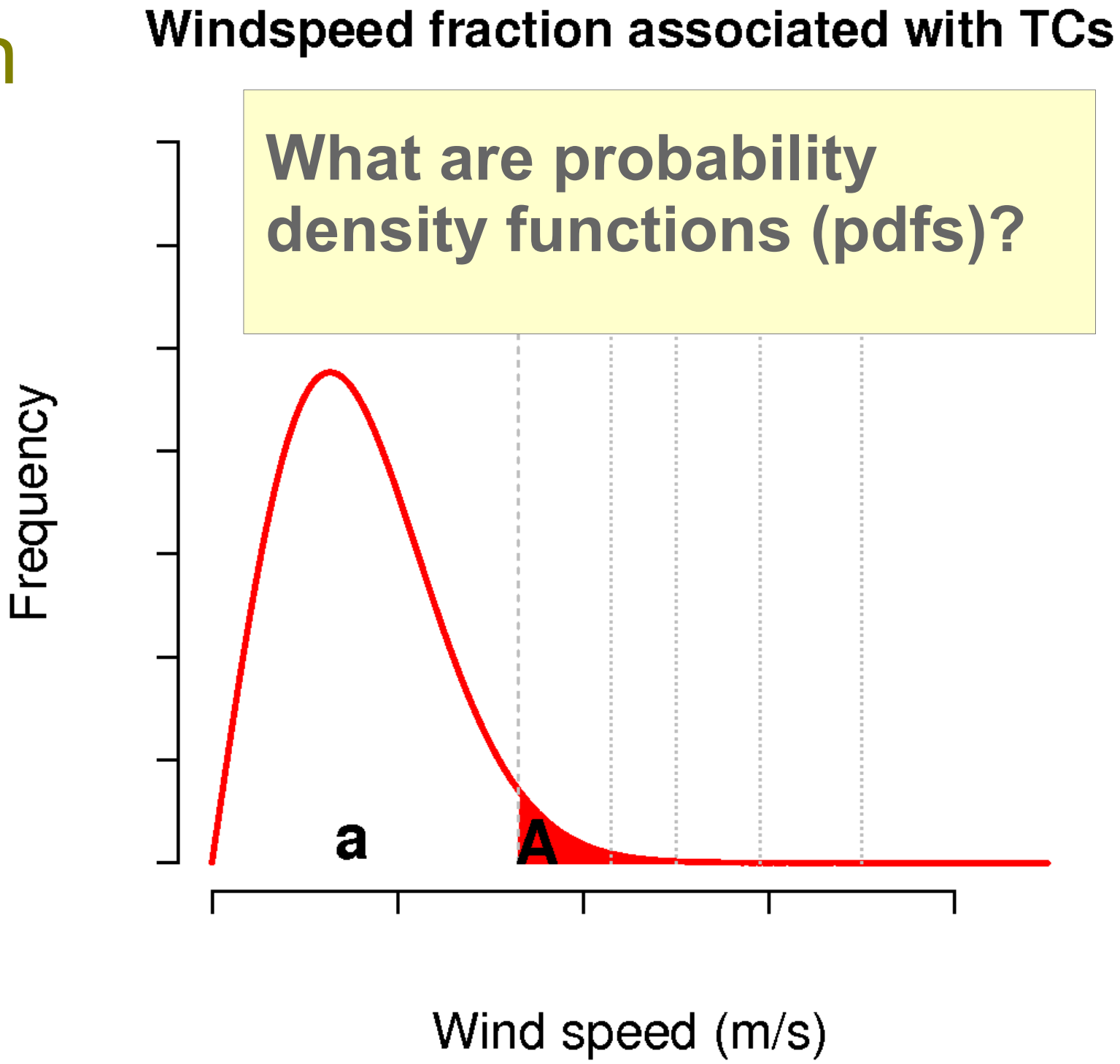
A measure of precision. "*Skill assessment is an objective measurement of how well the model nowcast or forecast guidance does when compared to observations*".[1]

Also a question of utility – how useful is the model?

If an incorrect model is useful, does it have skill?

http://www.nauticalcharts.noaa.gov/csdl/skillassess.html

Definition

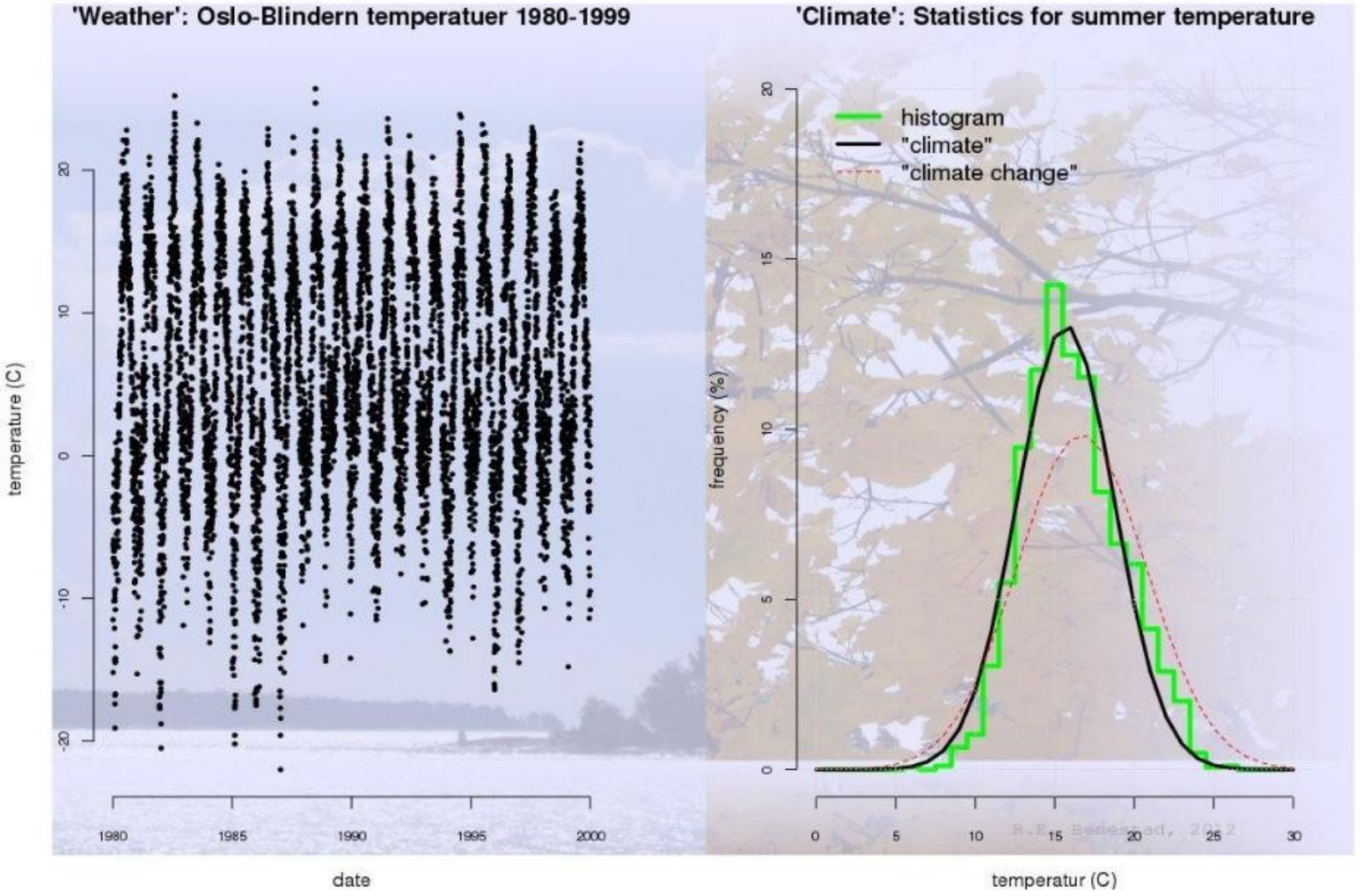

Windspeed fraction associated with TCs
What are probability density functions (pdfs)?
Frequency
a
A
Wind speed (m/s)

Frequency & Probabilities

Related: low probability → rare → low frequency: $p = f$

Example...

Weather (time series – chronological) & climate (pdf)




Skill

How do we measure skill?

- How closely a model describes the real world.
- Measure of reliability.
- Measure of precision.

Model skill – deterministic models

A **deterministic** model: y = g(x)

A deterministic model: a single number, completely specified by the inputs. Typically a weather forecast.

A 'good' model: y & g(x) are correlated – given by the equation.

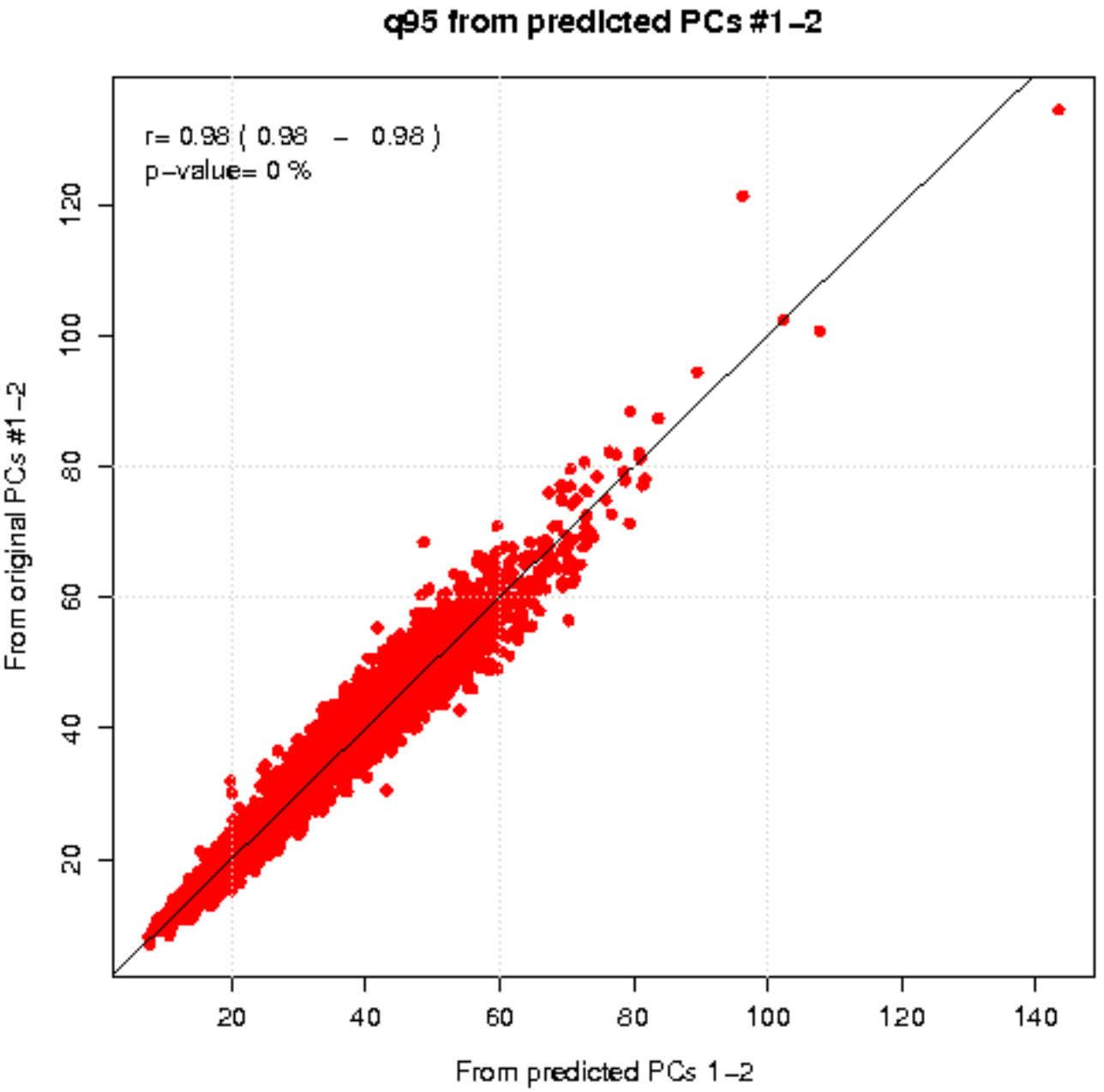
Common scores:

correlation, root-mean-squared-error (RMSE), contingency tables.

Correlation

- A verification of dependency
- X = Y?
- Scatter plots
- Pearson and ranked correlation.
- What correlation means dependency?

Scatter plots

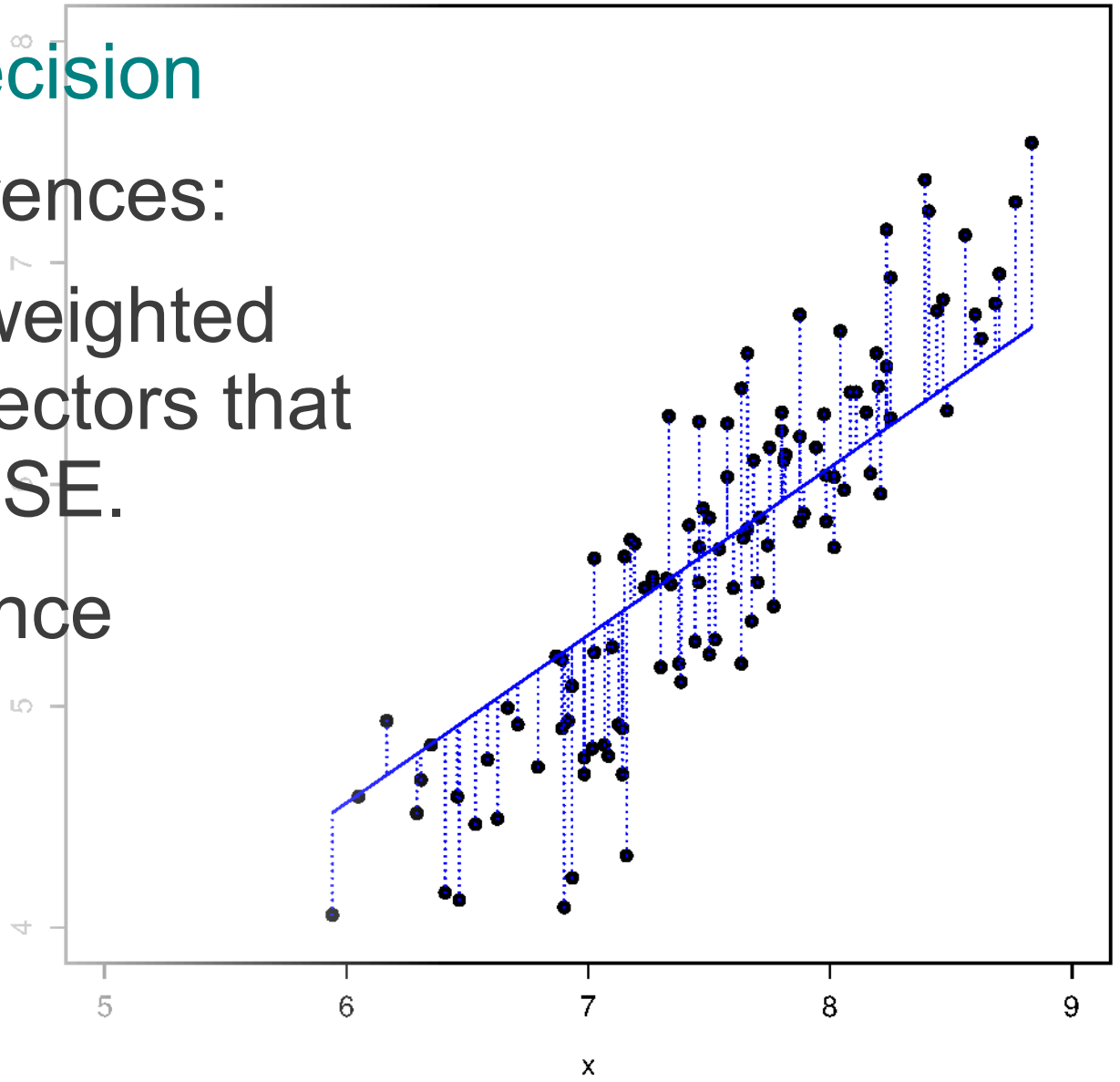


Corresponding values?

Graphical visualisation

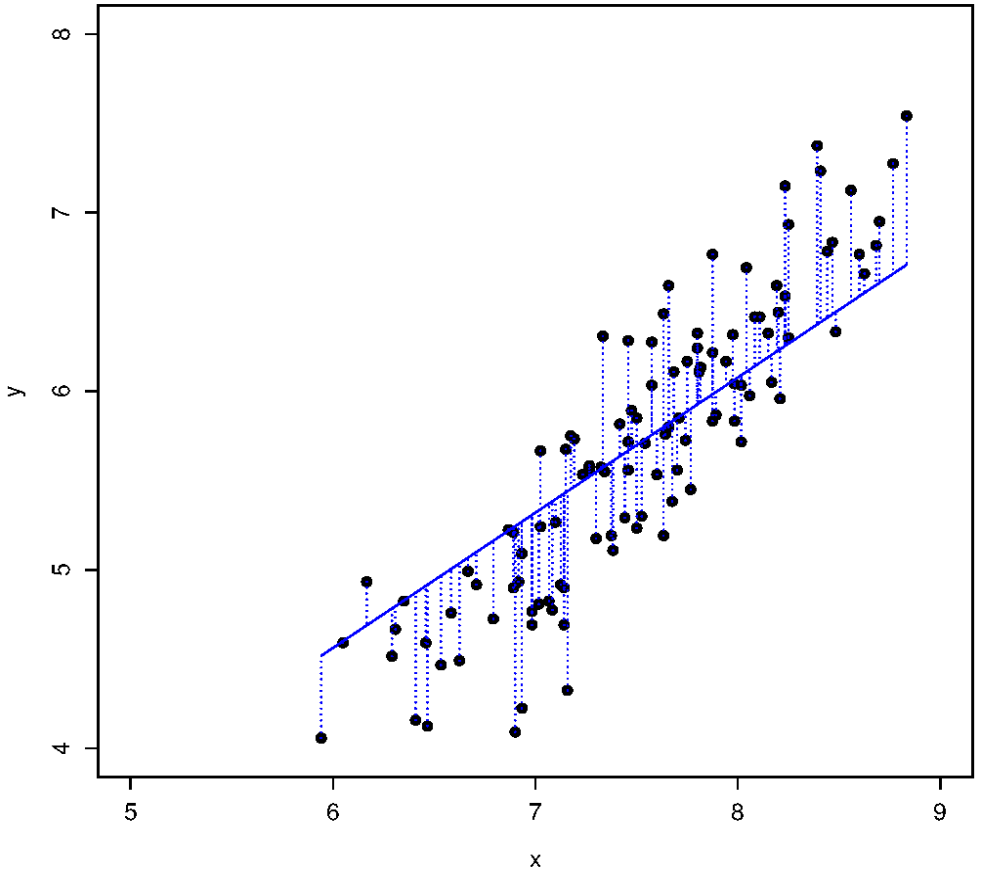
Root mean Square Error (RMSE)

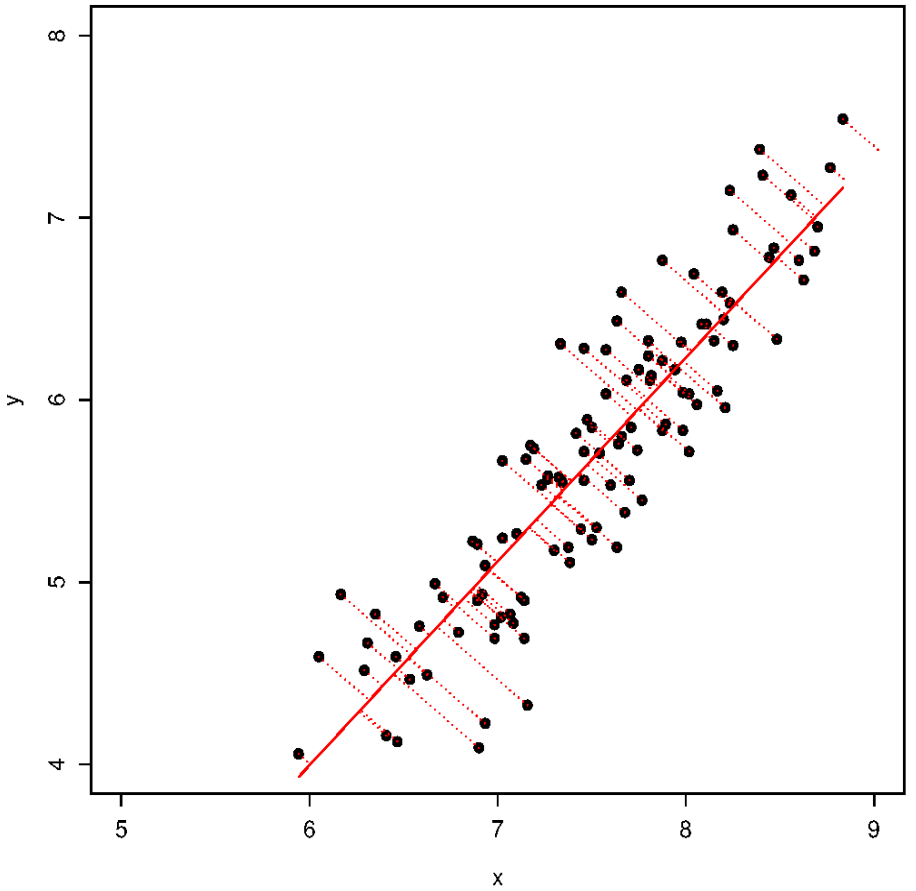
- Estimation of precision
- Emphasise differences:
- Regression – a weighted combination of vectors that minimize the RMSE.
- Analysis of variance (ANOVA).



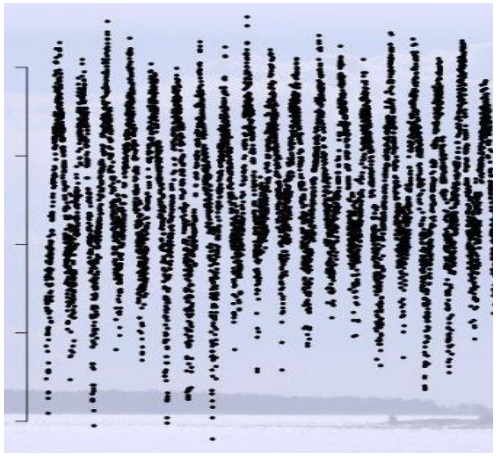
Least squares

2 types: minimizing the perpendicular distance to a line-fit and the errors in y:





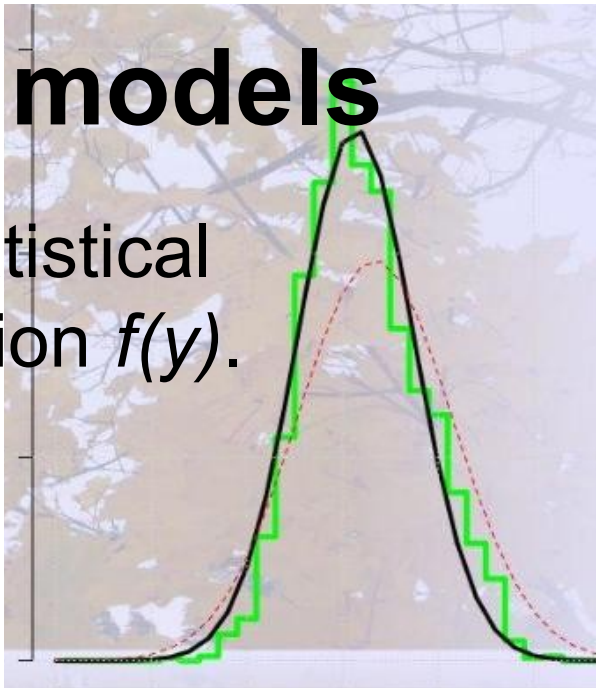
Statistical fingerprints for V&V.



- Correlations – dependencies.
- Time structure.
- Probability density functions (**pdfs**).

The physical system will leave a mark on the measured state. The *pdf* describes relative frequency. Correlations reveal dependencies. Cycles indicate the presence of constraints.

Model skill – probabilistic models



A **probabilistic** model: the output is a statistical description, in terms of spread & distribution *f(y)*.

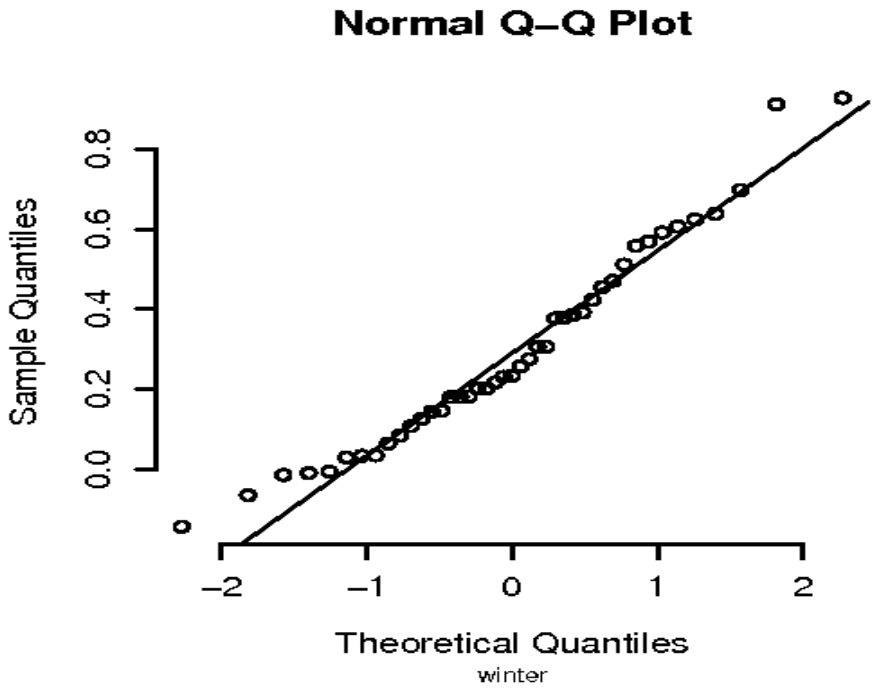
f(y) = g(x)

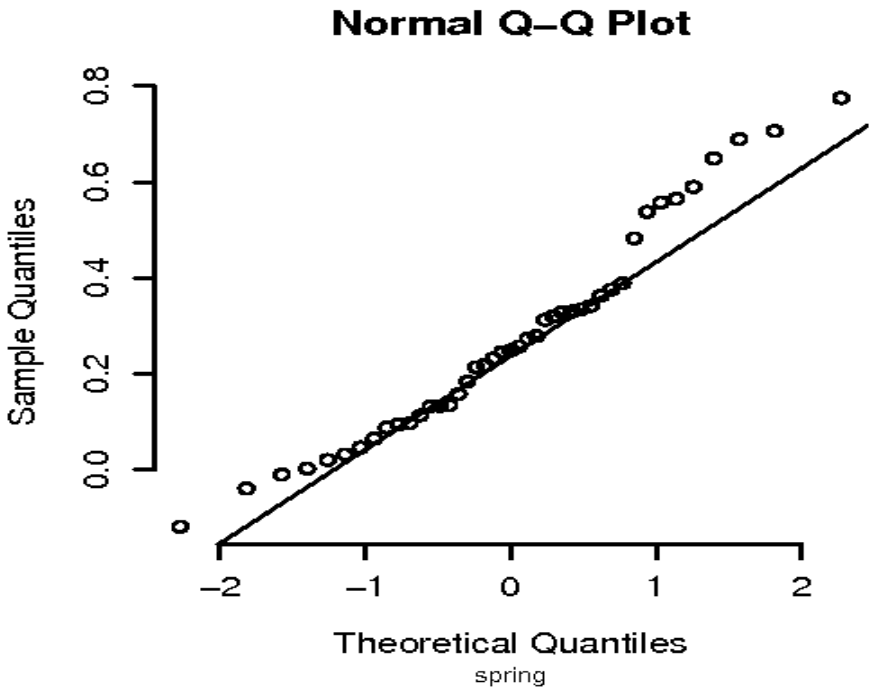
Climate predictions: *range & frequency*. Also, change in processes.

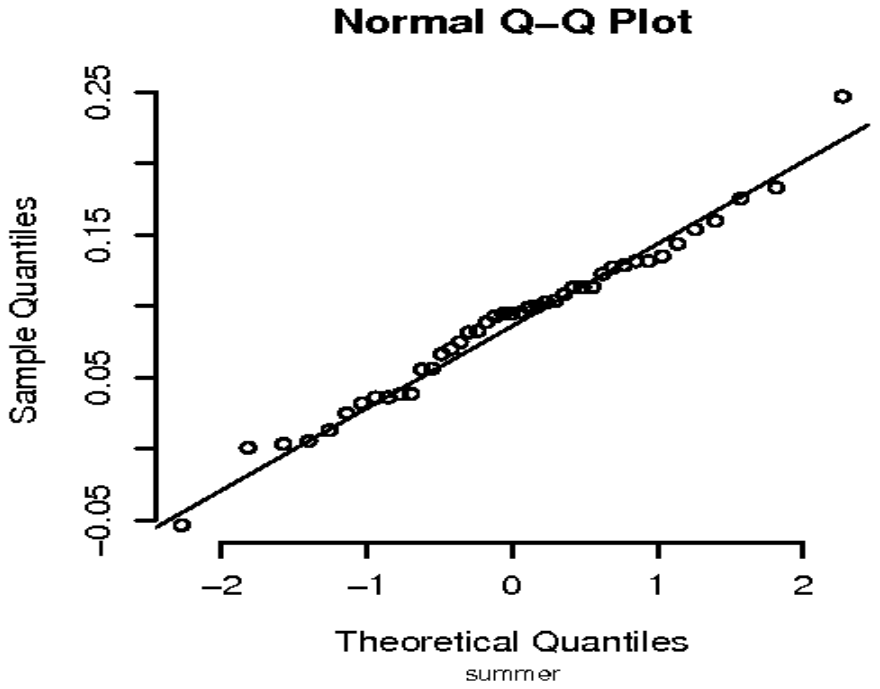
Skill scores:

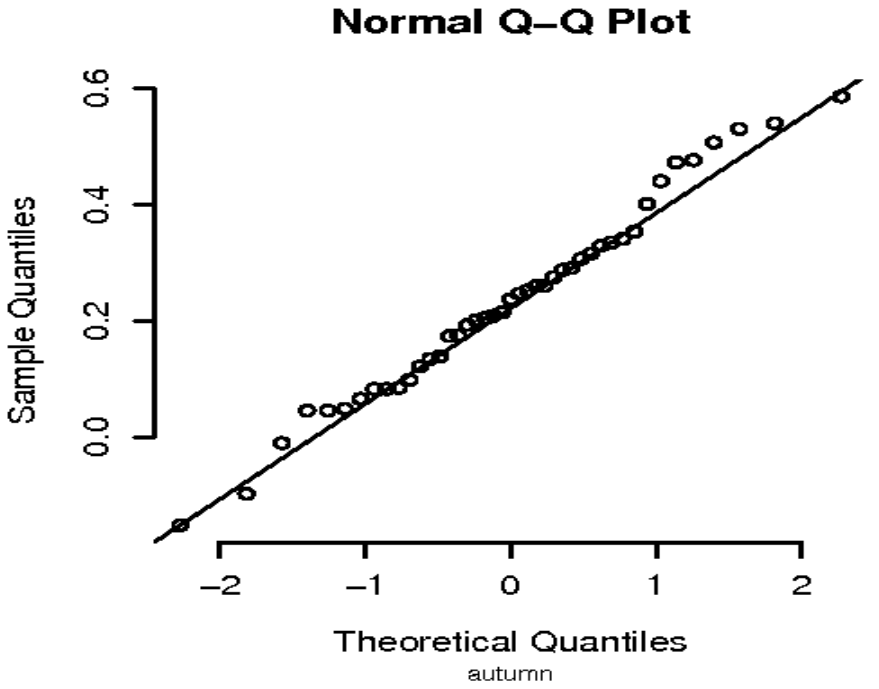
qq-plots, χ^2 (chi-squared), Student's t-test, Kolmogorov-Smirnov, Whitney-Mann U-test, Briers score, ROC-curves, Reliability diagrams, binomial distributions, Poisson distributions.

V&V: Are the distributions Gaussian?



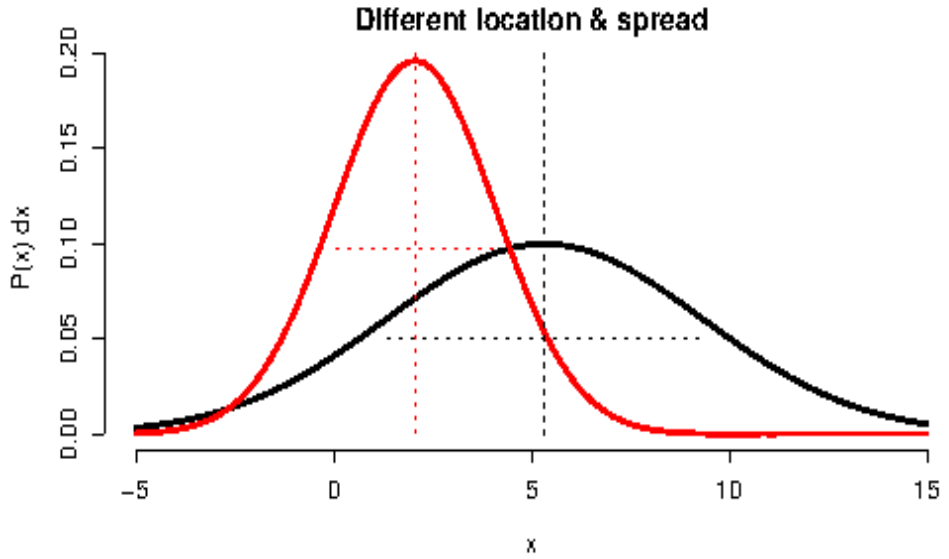






CMIP5 long-term trend coefficients

V&V: The student's t-test

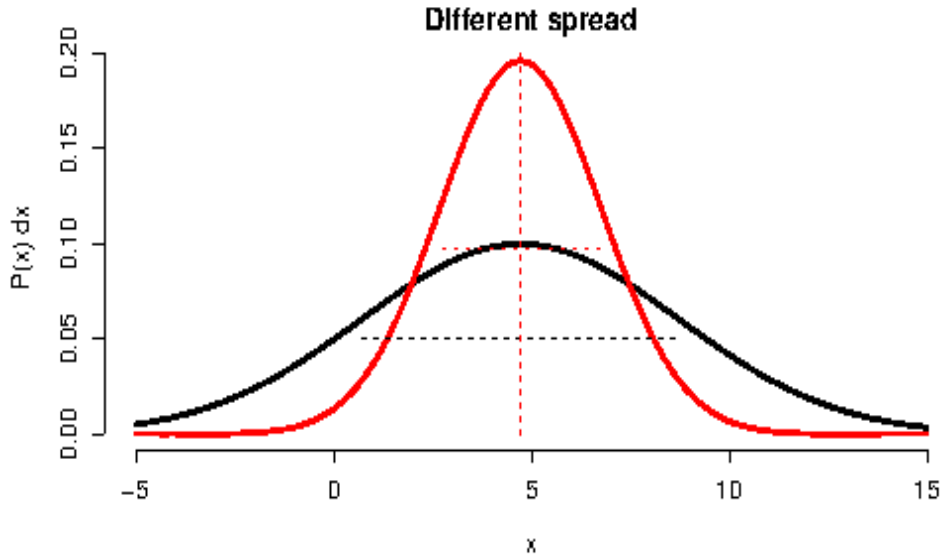


OK

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$$

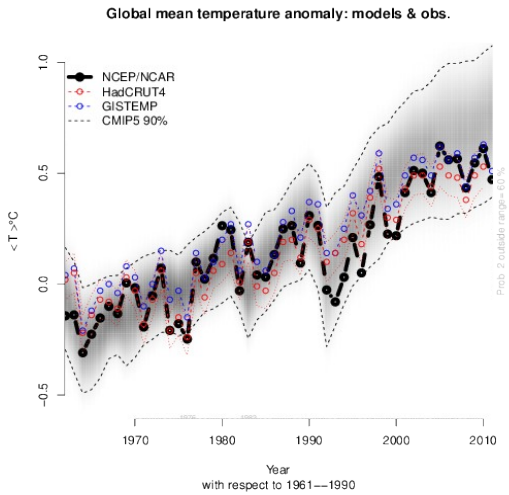


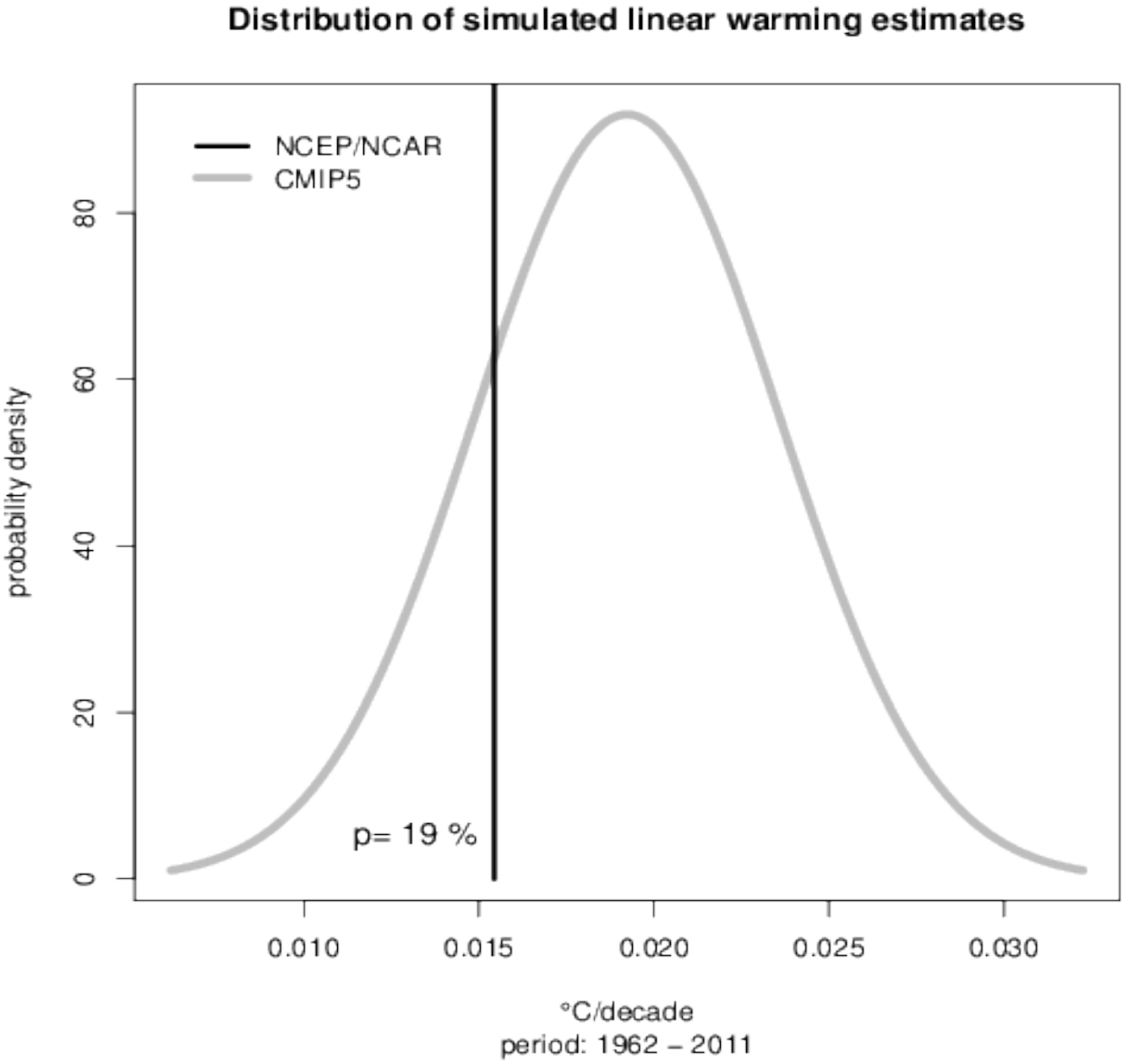
Fails

Bias and spread

- Often used in validating climate models
- Difference in mean
- Ignores a great deal of information
- Spread & Annual cycle
  - Simulate the processes well enough?

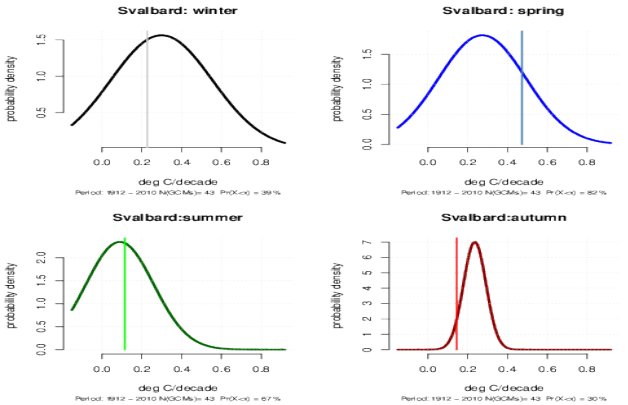
Question about trend

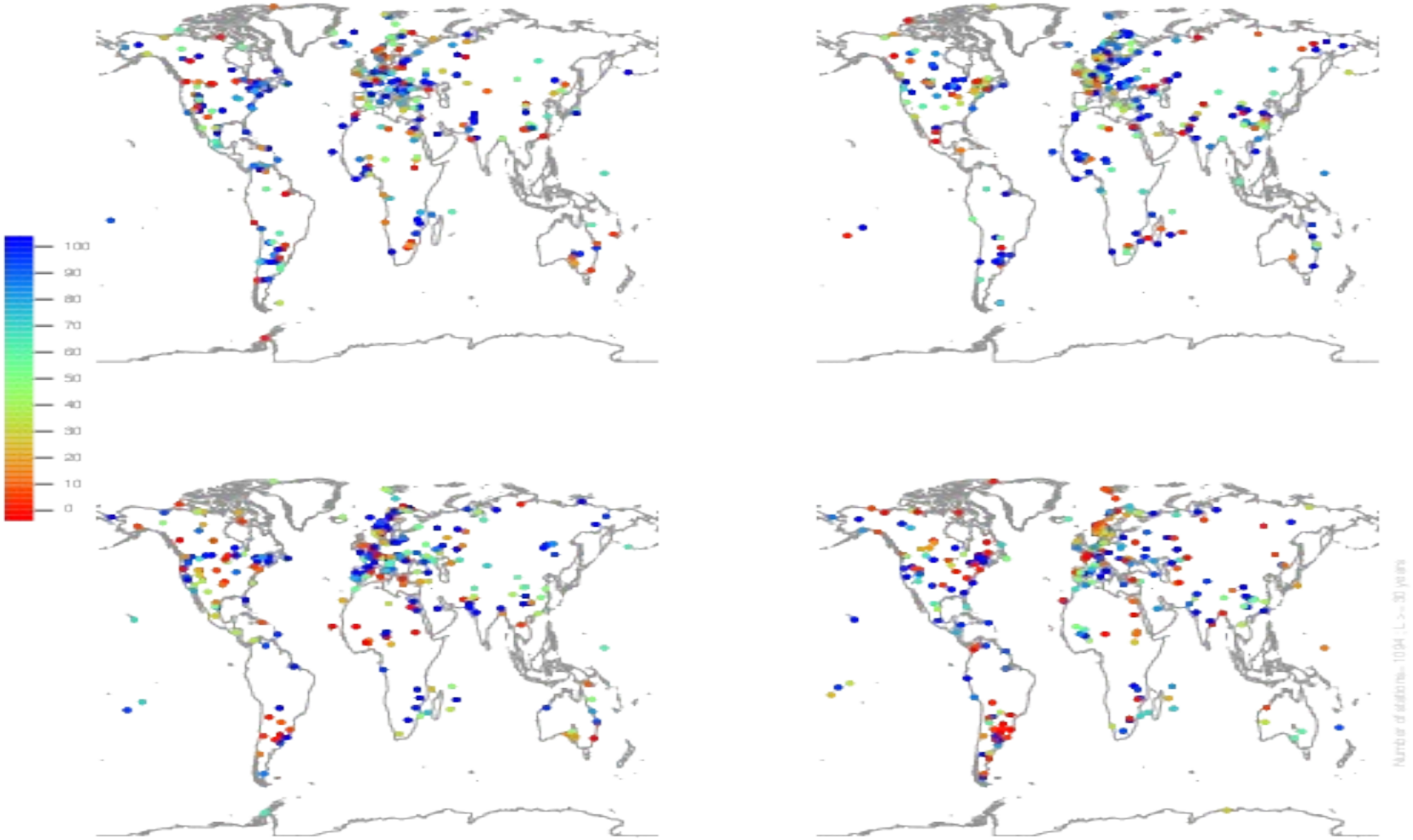




Predicting probabilities

- Rank verification





Contingency tables

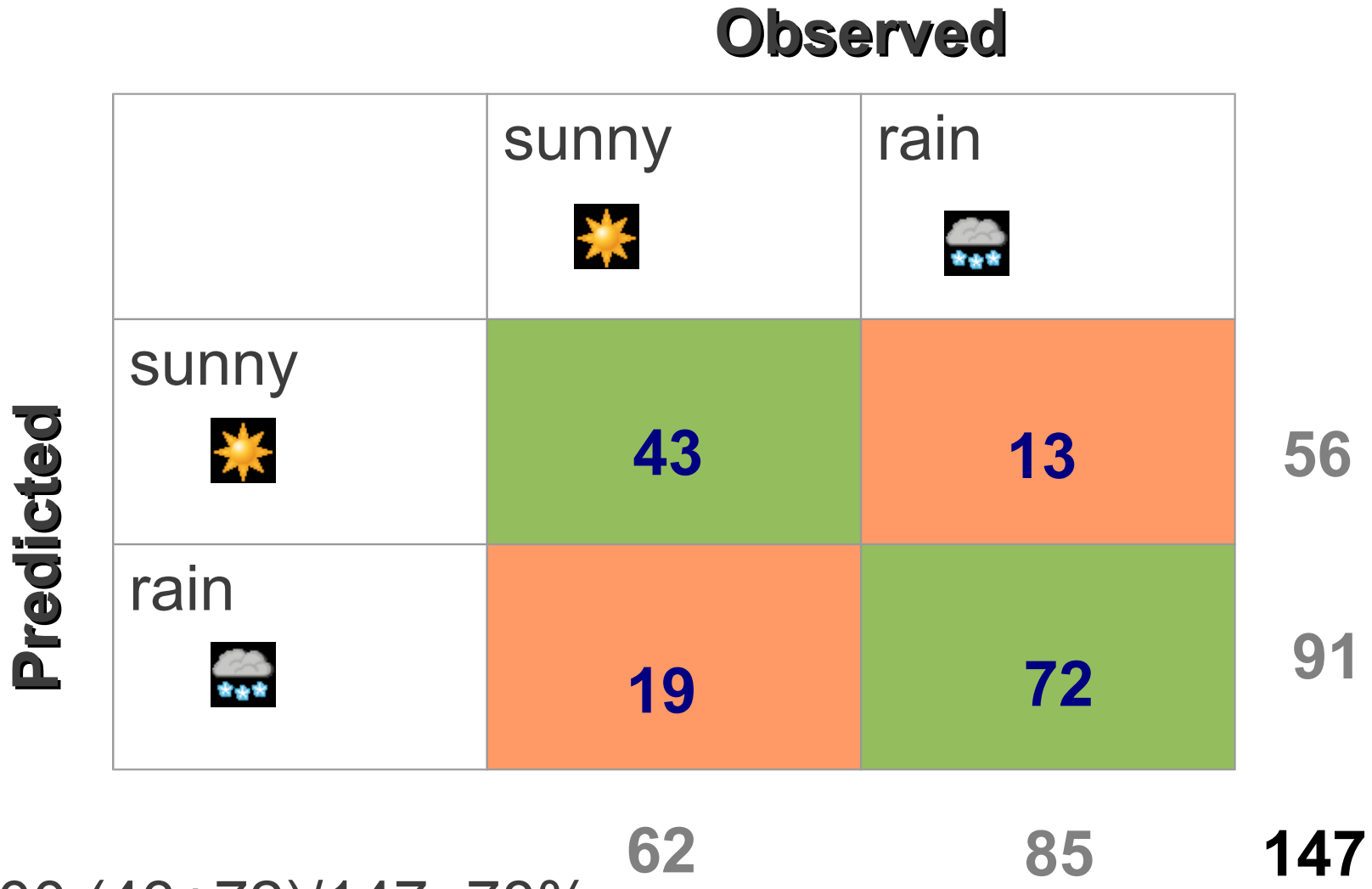
- Single quantities
- Different variables – different character
- Categorical predictions
  - Hit-ratio.
  - χ^2-test – a test of goodness of fit.

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed requency; E_i = expected frequency (probability), n=number of boxes in table.

Simple deterministic – contingency table

Hypothetical case: 147 forecasts..



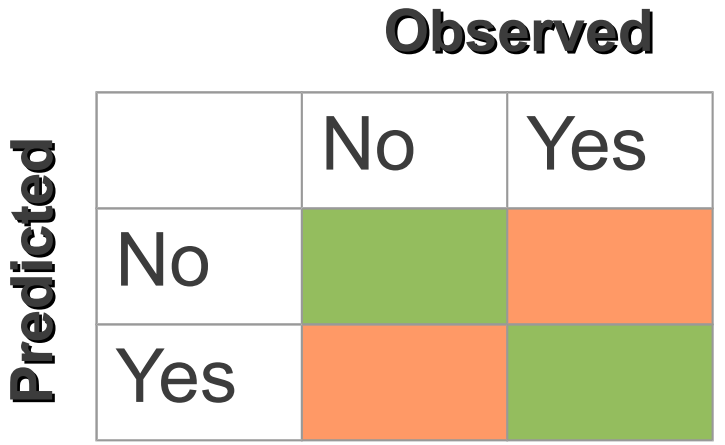
Hit ratio: 100 (43+72)/147=78%

Categorical forecasts

Finley's (1884) tornado forecast

- Very rare events do not make a mark on skill scores
- Higher scores by predicting “No” for all cases.
- More elaborate schemes to evaluate skill for extremes.

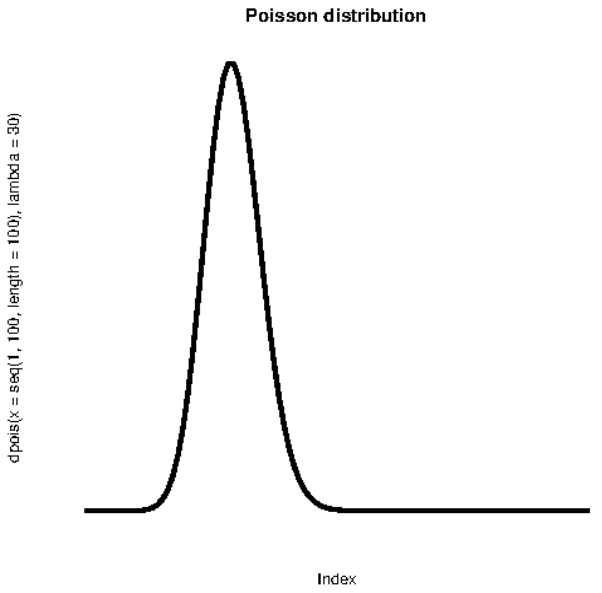
Binary forecasts – “*more subtle than they look*” (**Ian Joliffe**)

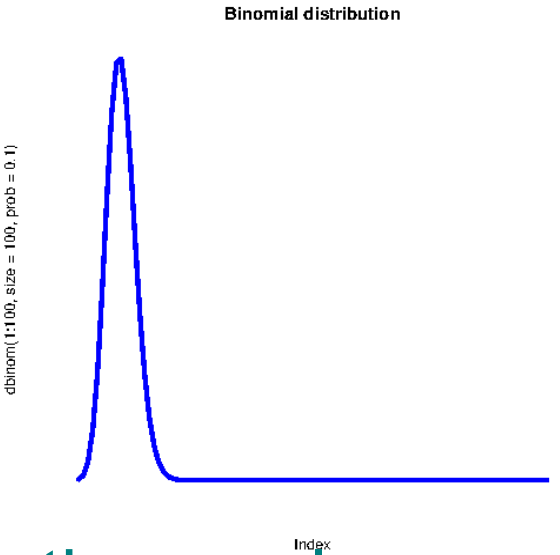


Predicting number of events

For random processes

- **Poisson and distributions:**
  - Number of cases, given a mean interval Λ between each.
  - $p(x) = \Lambda^x \exp(-\Lambda)/x! \quad X = [0,1,2,3,\ldots)$
- **Binomial distributions:**
  - Number of cases for a given *p* and sample size *n*.
  - $p(x) = \text{choose}(n,x)\ p^x\ (1-p)^{(n-x)}$



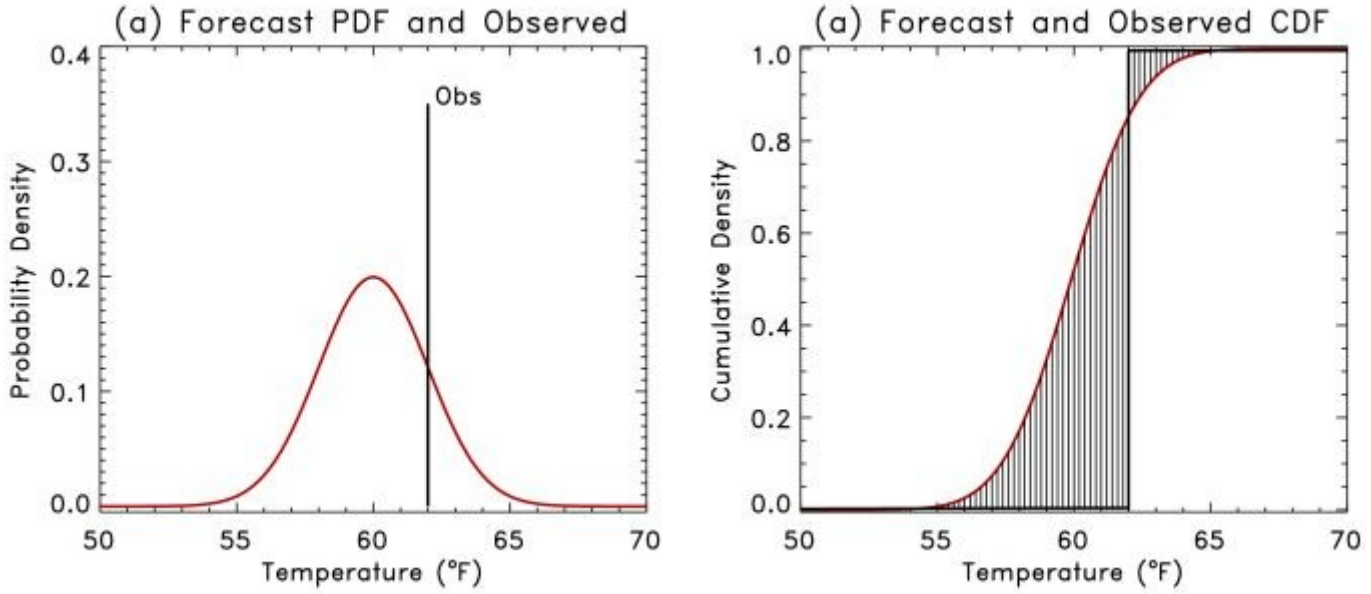


Deterministic processes will deviate from these rules

Predicting probabilities

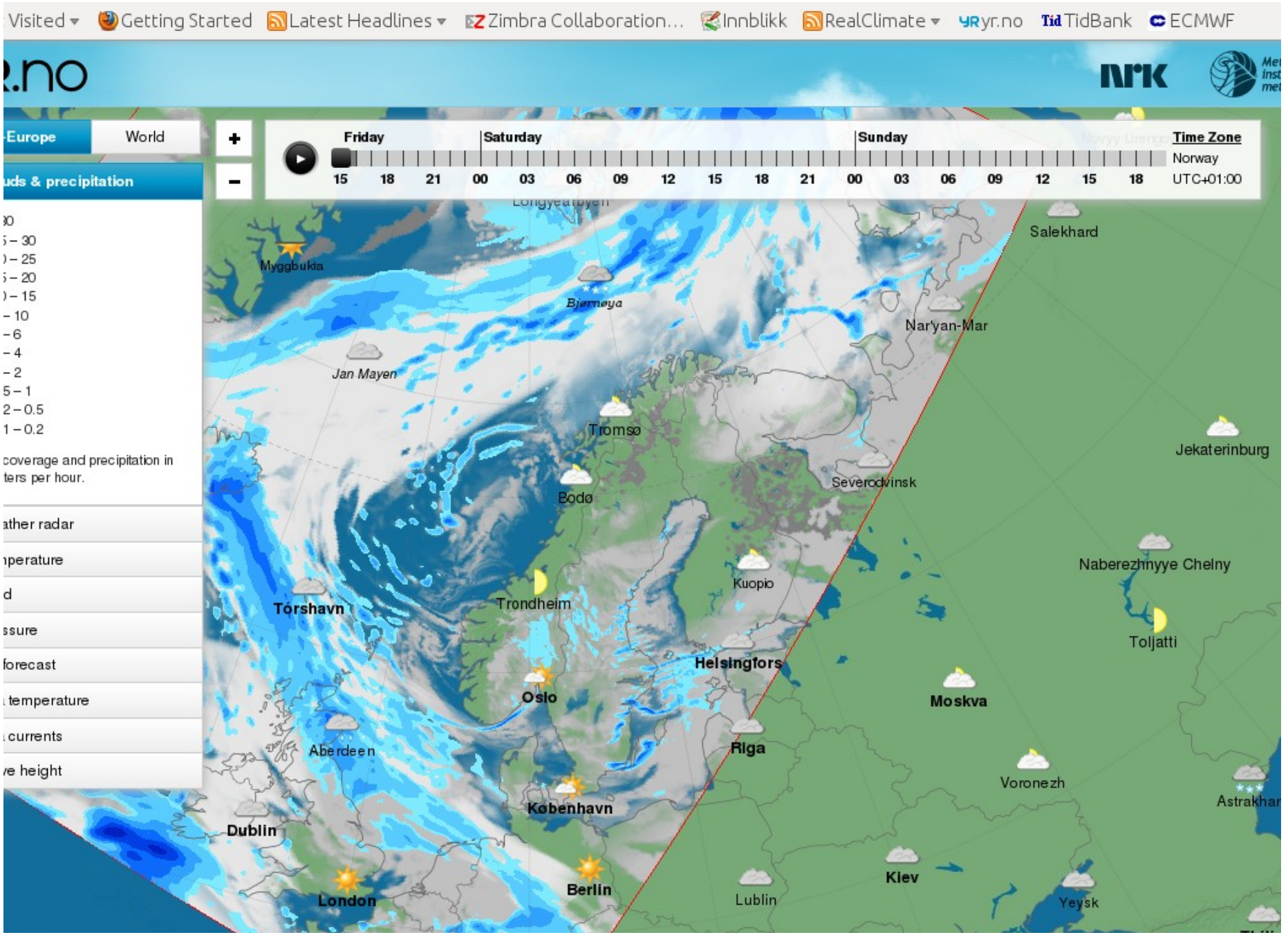
- Continuous ranked probability scores

$$CRPS(forecast) = \frac{1}{ncases} \sum_{i=1}^{ncases} \int_{x=-\infty}^{x=-\infty} \left(F_i^f(x) - F_i^o(x)\right)^2 dx$$



http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_prob_forec/uos3b/uos3b_ko1.htm

Weather forecasts & verification

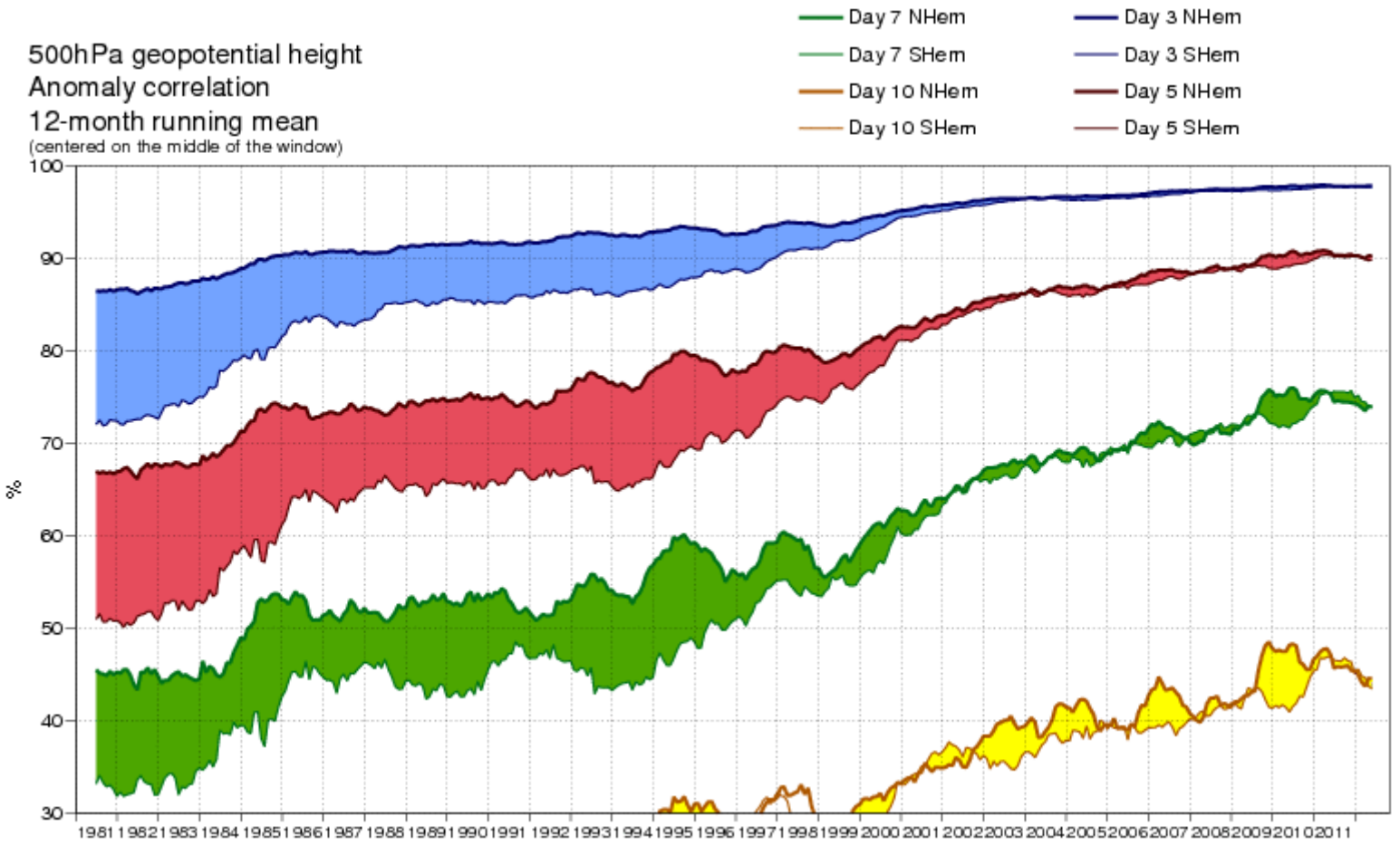


Weather forecasts – how to assess skill?

- Model simulations take current situation and compute the subsequent evolution.
- Atmospheric motion, temperatures, moisture, and phases (vapour or liquid).
- Time and space: right time or right place?
- Deterministic or probabilistic? How to evaluate predicted chances for rain?

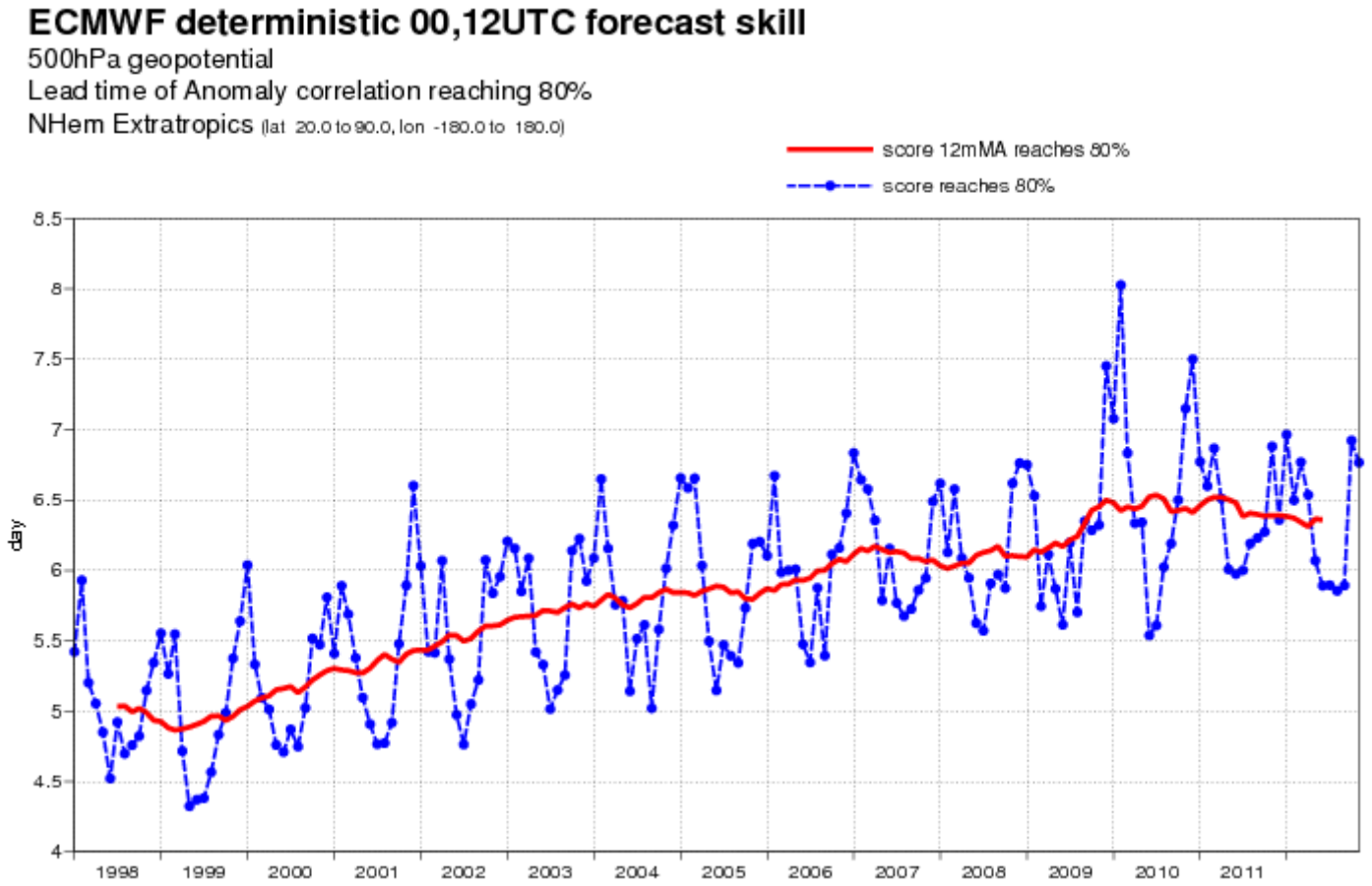
Weather forecast verification

- ECMWF
- Anomaly correlation of ECMWF 500hPa height forecasts



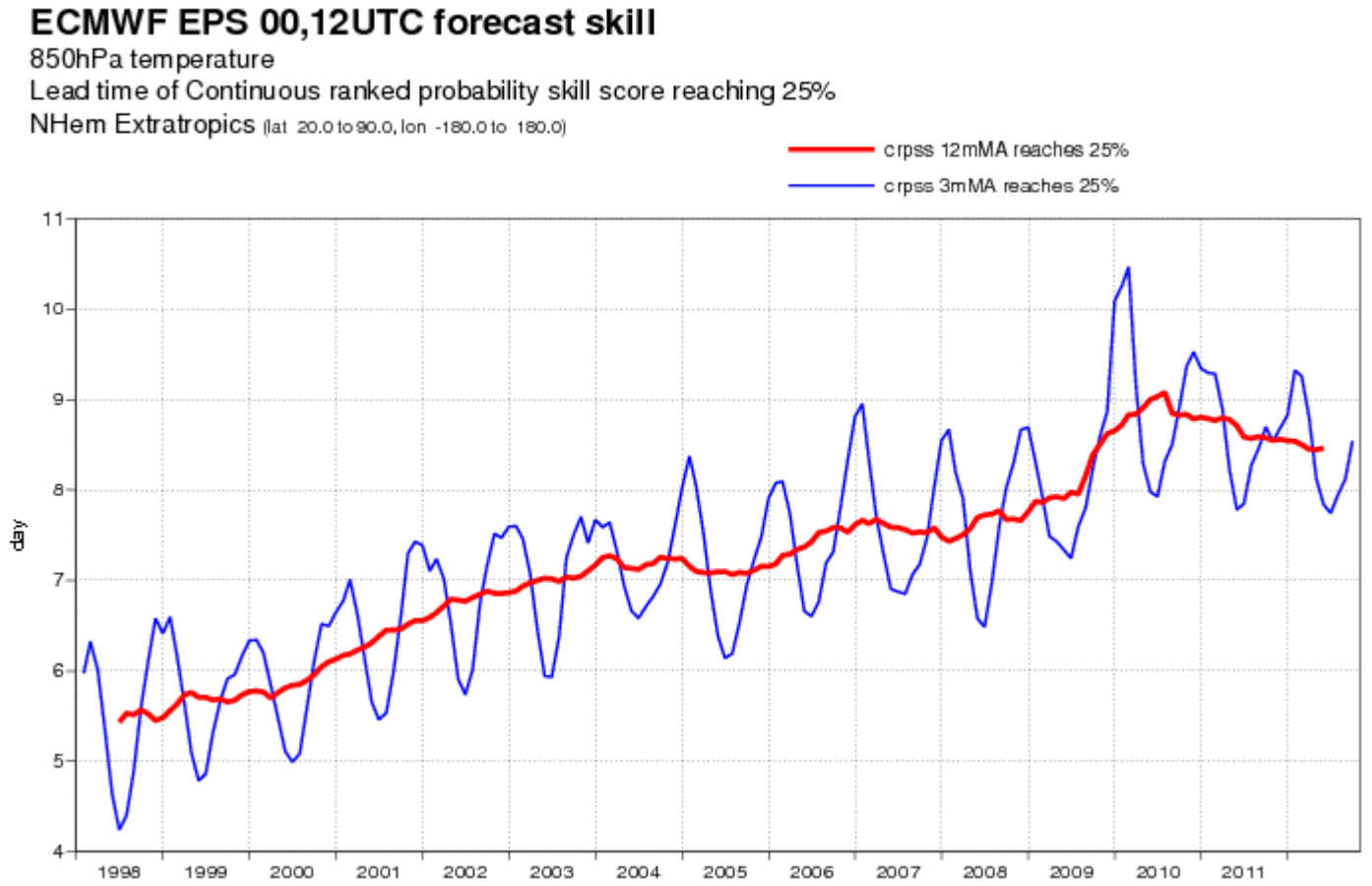
Deterministic forecasts

- Lead time – threshold score



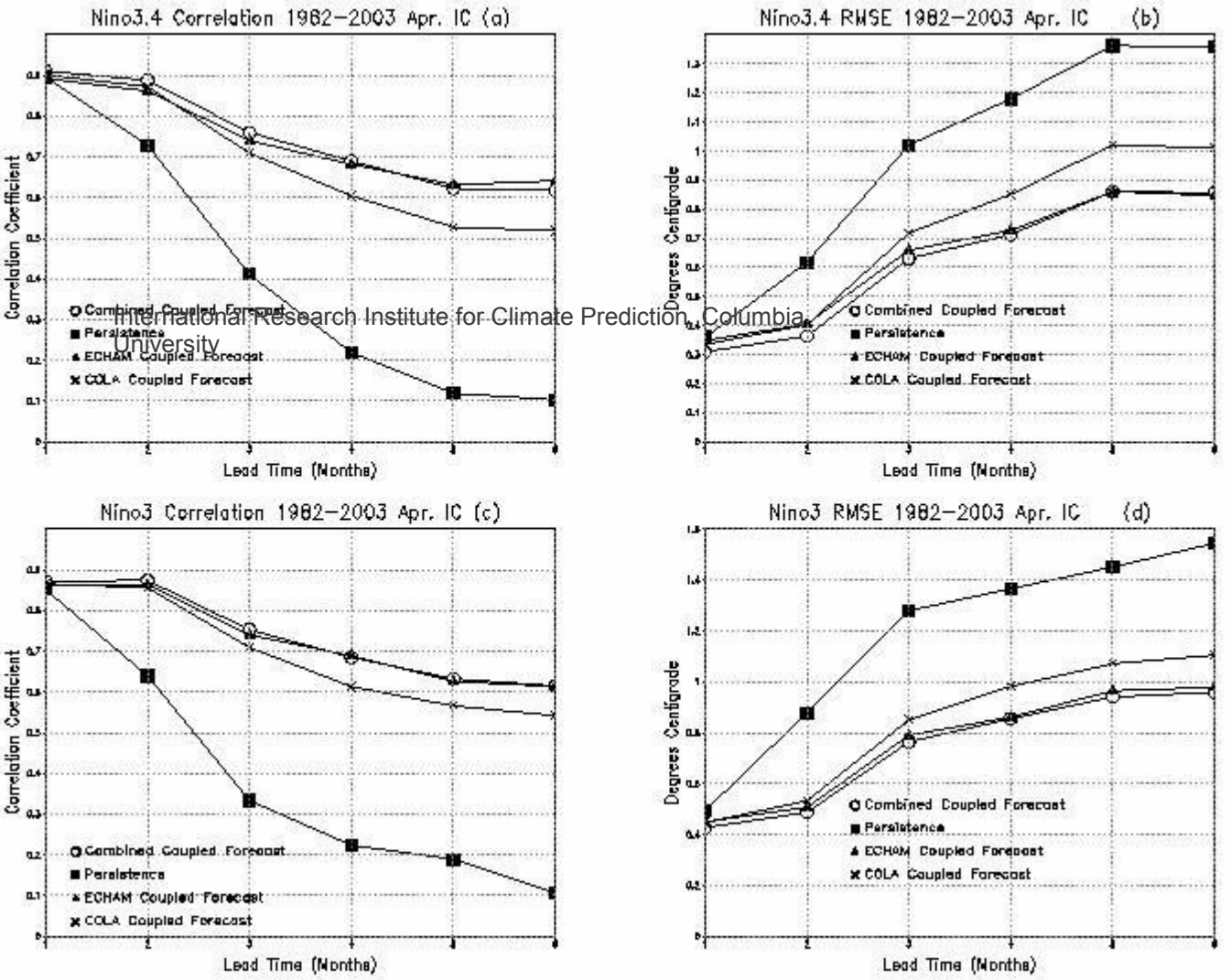
Ensemble forecasts

- Lead time – threshold score



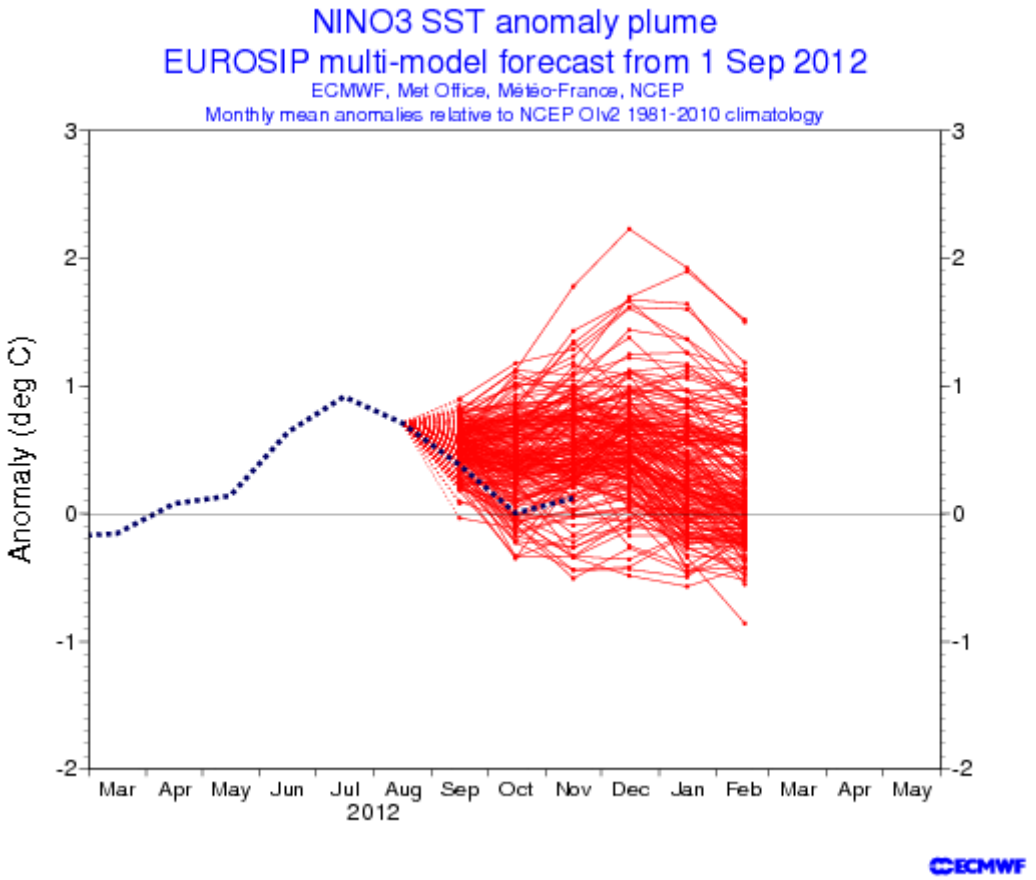
Nino3.4-index.

- International Research Institute for Climate Prediction, Columbia University



Seasonal forecasts

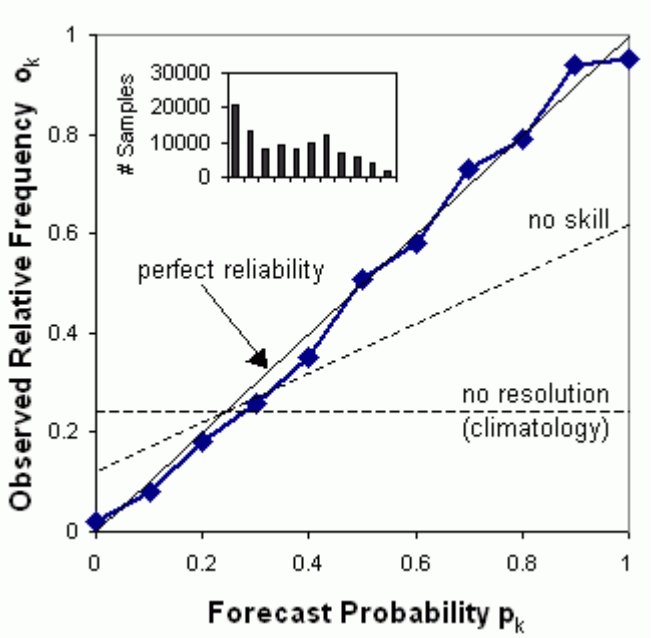
- 'Plume plot for ensemble forecasts



Reliability diagrams

- WWRP/WGNE Joint Working Group on Forecast Verification Research

http://www.metoffice.gov.uk/research/areas/seasonal-to-decadal/gpc-outlooks/user-guide/interpret-reliability



The Brier score:

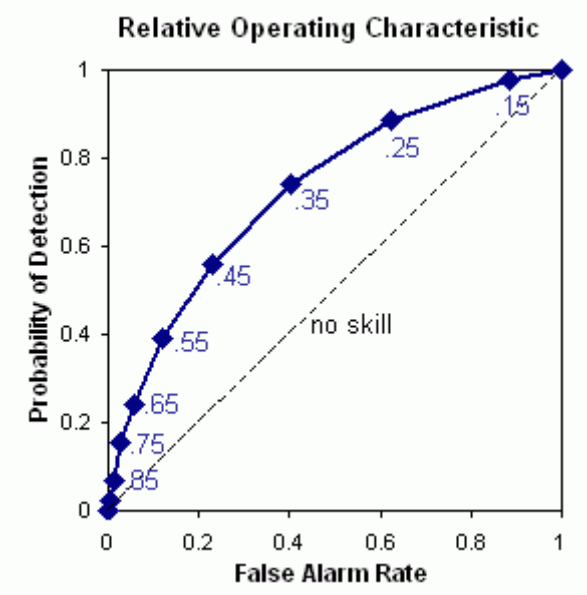
$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

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

http://www.metoffice.gov.uk/media/pdf/j/6/SVSLRF.pdf

Relative operating characteristic

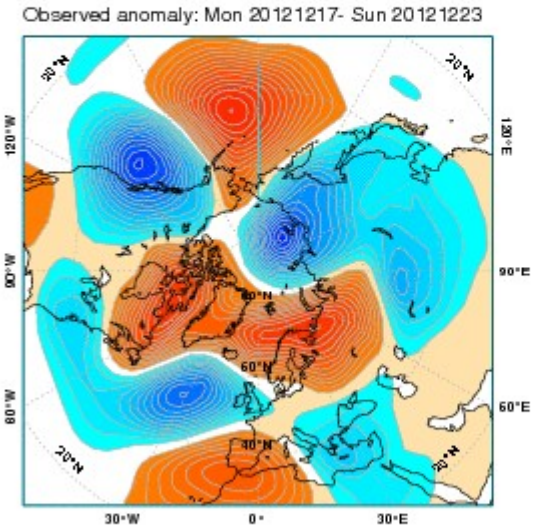
- WWRP/WGNE Joint Working Group on Forecast Verification Research

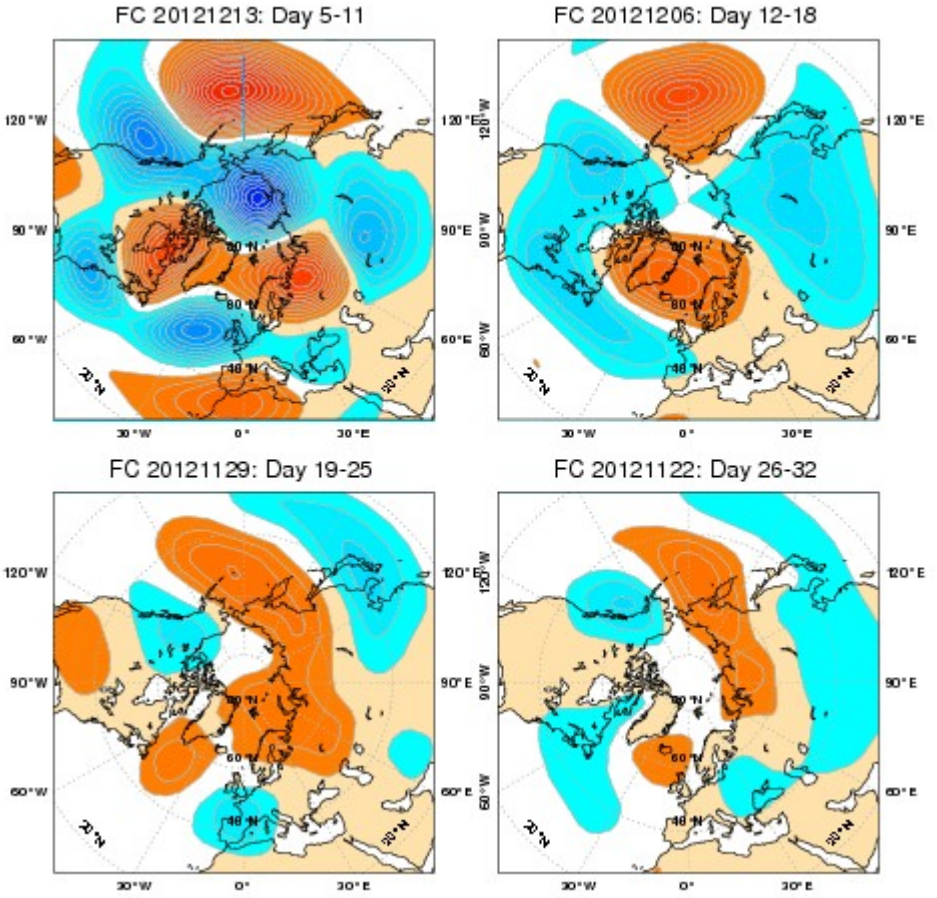


http://www.cawcr.gov.au/projects/verification/verif_web_page.html

Monthly forecasts

- Maps of anomalies.
- Spatial correlations





Next lecture