Validation for extreme events

Joanna Wibig
Department of Meteorology and Climatology
University of Lodz

Validating Regional Climate Projections,
4th VALUE Training School – Triest, 26-30 Oct 2015

Sources of the errors and uncertainties of downscaled climate simulations

- ***an imperfect model formulation,***
  - ***errors of the driving GCM,***
  - ***errors inherent in the downscaling approach,***
  - ***errors in observations themselves,***
- ***uncertain future concentrations of GHGs,***
- ***internally generated climate variability.***

Validation – what does it mean?

Assurance of the quality of the projection in one specific aspect

We do not treat validation as synonym of verification.

Validation needs observation data

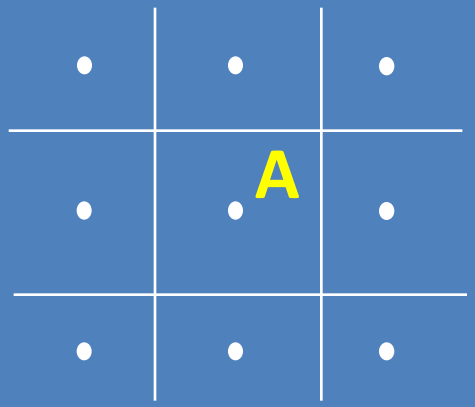
- ❑ *reference data set,*
- ❑ *forcing in a perfect boundary setting*

Quality of data

- *inhomogenities,*
- *outliers,*
- *biases*

Availability of long reference data sets

Station data versus point data

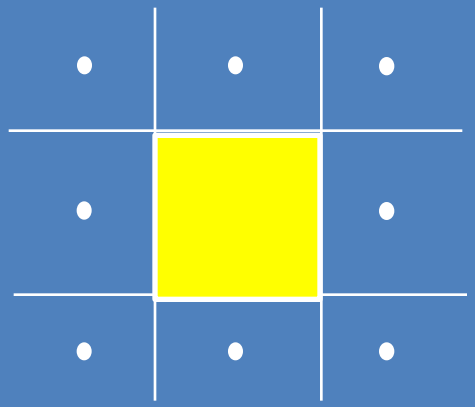


Observation:

a convective rain of 60mm/day , in a surroundings of point A (on the area of 10 km²)

Availability of long reference data sets

Station data versus point data



Observation:

a convective rain of 60mm/day , in a souroundings of point A (on the area of 10 km²)

Projections:

a convective rain of 6mm/day, in a grid containg point A (area of 10 x 10 = 100 km²)

Gridded data sets obtained from observation/station data

- *interpolation*
- *averaging*
- *density of station data*
- *type of data – spatial variability*
- *effect on extremes*

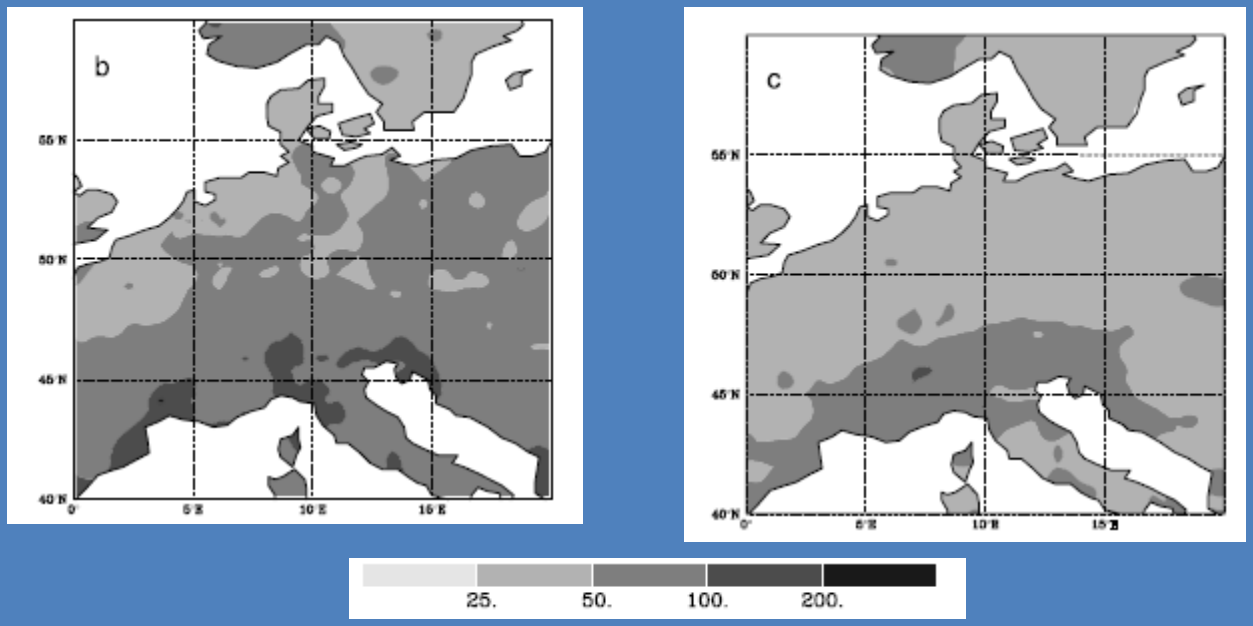
E-OBS daily data set of temperature and precipitation (Haylock et al. 2008) the ECA&D (http://eca.knmi.nl; Klok and Klein Tank 2009)

Haylock et al. (2008) mention that one of the main reasons affecting the behaviour of extremes in gridded data is related to the interpolation methodology, as non exact interpolators smooth out peaks and troughs in a surface.

The median reduction factor for the 10-year return level is 0.66 for precipitation and more than 1 degree C for maximum temperature, but it can be even below 0.5 for precipitation and more than 3 degrees C for maximum temperature as compared with raw station data (Haylock et al., 2008).

The effect of gridding on extremes

Haylock et al., 2008



Ten-year return level for daily precipitation for (b) station data, and (c) the final gridded data.

Reanalysis as a substitute for observation data

- ***numerical model hindcasts into which observational data have been assimilated***
- ***it is necessary to check whether the observations representing the variable of interest have been assimilated into the model***

- *NCEP/NCAR Reanalysis (Kalnay et al. 1996)*
- *ERA-40 (Uppala et al. 2005)*
- *ERA-Interim (Dee et al. 2011)*
- *North American Regional Reanalysis (Mesinger et al. 2006)*

Classification of the reanalysis gridded fields

- ***The most reliable class. Variable is strongly influenced by observations (e.g. upper-air temperature);***
- ***Less reliable. Observations directly affect the variable, however also model has a strong influence (e.g.humidity or surface temperature);***
- ***Observations do not affect the variable, so it is solely derived from the model (e.g. clouds, precipitation, surface fluxes).***

Kalnay et al. 1996

INDICES

Indices should depend on the validated aspect of projected climate

- ***Application of the downscaled product***
- ***Temporal scale,***
- ***Spatial scale***

Indices need to be independent of calibration and tuning (calibration and validation should be done on different data sets)

INDICES

Comprehensive lists of indices are available from:

the 'Expert Team on Climate Change Detection and Indices' (Peterson et al. 2001),

http://etccdi.pacificclimate.org/docs/ETCCDMIndicesComparison1.pdf

the STARDEX project (Goodess et al. 2005),

http://www.cru.uea.ac.uk/projects/stardex/

the ENSEMBLES project (van der Linden and Mitchell 2009).

INDICES

To validate the distribution of the variable of interest:

- ***mean***
- ***variance***
- ***specific quantiles (for instance, for strong but not yet extreme events - 90th percentile).***

More generally, the indices can be the parameters of a parameterised formulation of the distribution.

INDICES

The representation of extreme events can be based on parametric distributions build on:

- ***the extreme value theory, the generalised extreme value (GEV) distribution, to validate maxima of long blocks***
- ***the generalised Pareto distribution (GPD), to validate excesses of high thresholds.***

PERCENTILES

- ***R95q – highly wet day percentile***
- ***R95 – number of highly wet days***
- ***R95p – percentage of precipitation on highly wet***
- ***R95t – total precipitation on highly wet days***
- ***R95m – mean daily total on highly wet days***

INDICES

- ***Return levels: 10 - 20 year, 100-year***
- ***Threshold exceedance***
- ***1,3,5-days highest totals***

INDICES

Temporal indices used for weather extremes:

- ***the length of extreme events such as droughts or wet spells,***
- ***the maximum length of an event in a defined period, such as a season.***

Validate using the data directly with grid box resolution, or smooth the data in advance?

- *regional climate simulations are not meant to be interpreted on a grid box level and so the former choice would be too rigid.*
- *RCM simulations are often used on the grid box level, and a validation should not influence the corresponding performance.*
- *In impact studies, the simulated unsmoothed fields are often required even when they are not interpreted on a grid box level and smoothing might hide important spatial properties such as the spatial correlation structure.*

MEASURES

They are used to quantify the discrepancy between the modelled and reference validation

Statistical tests which explicitly address the significance of the discrepancies:

- ***False positive results***
- ***Too low power of a test to detect model errors due to a lack of data (short records)***
- ***A significant deviation might simply be completely irrelevant.***

In some validation studies, the discrepancies have only been visually inspected.

Distributions vs. event – wise validation

In a control run setting there is lack of relation between the weather sequences and model simulation sequences. The validation can therefore be based only on long-term statistics - distributions. Climates not weather are compared.

Distributions vs. event – wise validation

In a control run setting there is lack of relation between the weather sequences and model simulation sequences. The validation can therefore be based only on long-term statistics - distributions. Climates not w[illegible]er are compared.

DISTRIBUTION-WISE

Distributions vs. event – wise validation

In a perfect boundary setting, we suspect that the modelled and observed weather sequences are more or less synchronous. The validation can therefore, be based not only on long-term statistics (distribution-wise validation), but also the measures developed for of weather forecasts can be applied (event-wise validation). Weather events can also be compared.

Distributions vs. event – wise validation

In a perfect boundary setting, we suspect that the modelled and observed weather sequences are more or less synchronous. The validation can therefore, be based not only on long-term statistics (distribution-wise validation), but also the measures developed for of weather forecasts can be applied (event-wise vali n). Weather events can also be compa

EVENT-WISE

MEASURES

Typical measures of the similarity of distributions of weather variables are discrepancies in mean and variance. Generally t-test is used for equality of means and the F-test for equality of variances.

- ***Both tests relay on strong distributional assumption (normal distribution)***
- ***T-test is unable to detect changes in scale***
- ***F-test is unable to detect changes in location***
- ***The power of both tests lowers significantly if the distributions are not normal***

Wilks 1995, Wilcox 1997, Ferro et al., 2005

MEASURES

Three resistant measures of the location, scale and shape (assymetry) of distribution function F were proposed by Lanzante (1996, IJoC)

Ferro et al., 2005. IJoC

$$m_X = \hat{x}_{0.5}$$ *median*

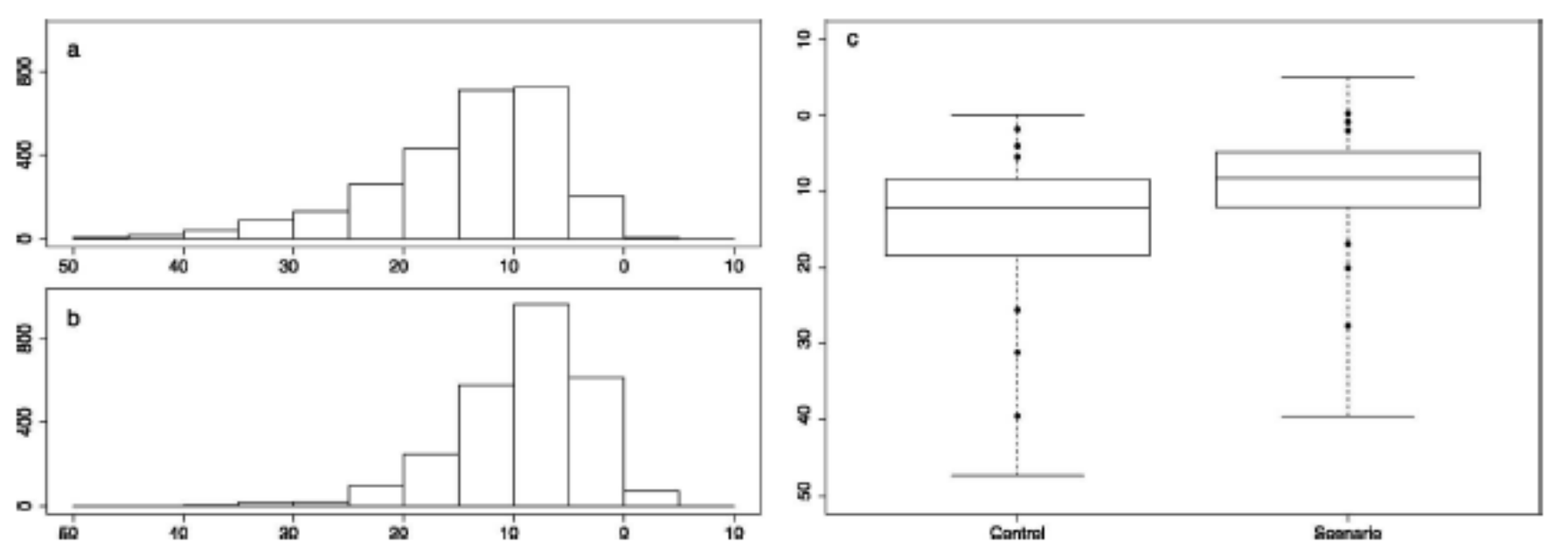
$$s_X = \hat{x}_{0.75} - \hat{x}_{0.25}$$ *interquartile range*

$$a_X = (\hat{x}_{0.75} - 2\hat{x}_{0.5} + \hat{x}_{0.25}) / s_X$$

Yule-Kendall skewness measure

Ferro et al. , 2005. IJoC

MEASURES



Histograms of (a) control and (b) scenario DJF daily minimum temperatures (°C); (c) boxplots of control and scenario temperatures. The boxplot whiskers extend over the range of the data; three lower quantiles (*p* 0.01, 0.05, and 0.1) and three upper quantiles (*p* 0.9, 0.95, and 0.99) are marked (●).

MEASURES

the overall distribution comparison:

- ***the chi-square test,***
- ***the Kolmogorov–Smirnov test***
- ***quantile (QQ) plots***

Ferro et al., 2005. IJoC

Hypotheses:

$$H_0 : F(z) = G(z)$$

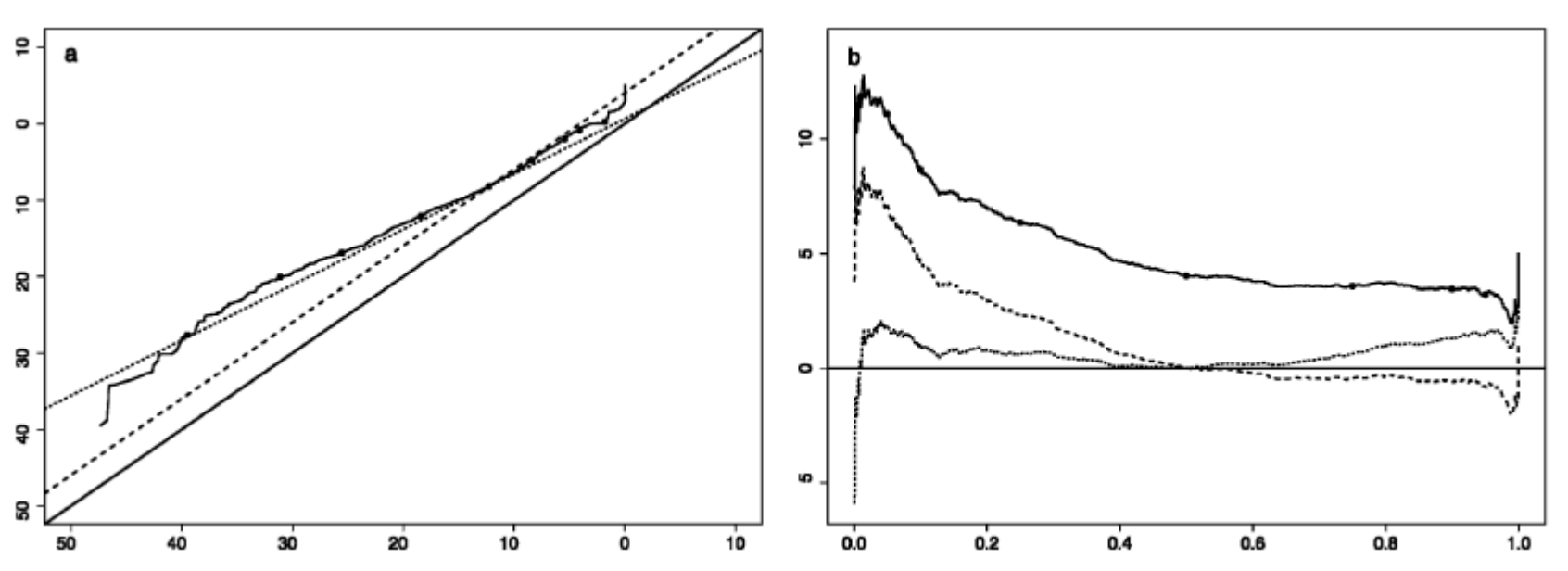
$$H_S : F(\sigma_x \cdot z) = G(\sigma_y \cdot z)$$

$$H_L : F(\mu_x + z) = G(\mu_y + z)$$

$$H_{LS} : F(\mu_x + \sigma_x \cdot z) = G(\mu_y + \sigma_y \cdot z)$$

MEASURES

Ferro et al. , 2005. IJoC

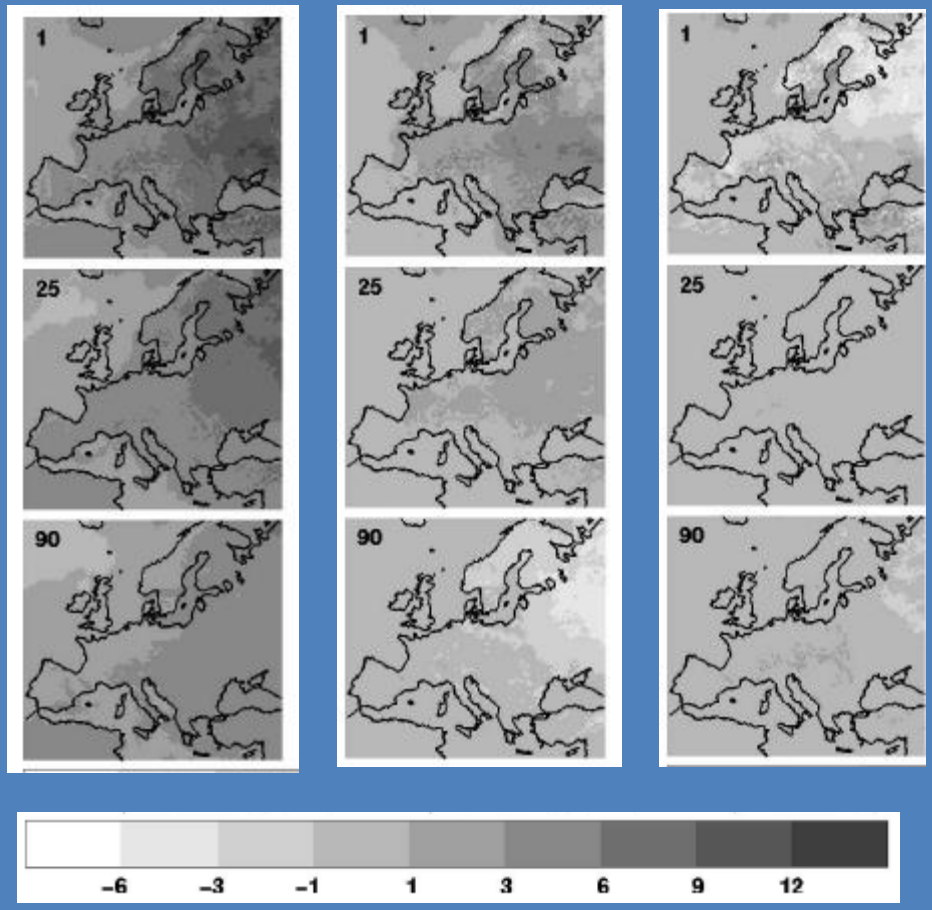


(a) Quantile–quantile plot (line with dots) of scenario vs control DJF daily minimum temperatures (°C) with straight lines corresponding to hypotheses *H*0 (normal line), *HL* (dashed line), and *HLS* (dotted line); (b) quantile differences (line with dots), location adjusted (dashed line), and location and scale adjusted (dotted line) against probability.

MEASURES

Differences (°C) in selected quantiles of DJF daily minimum temperatures between two distributions

raw	*adjusted for location*	*adjusted for location & scale*



It is possible to analyse which parameter of distribution has changed

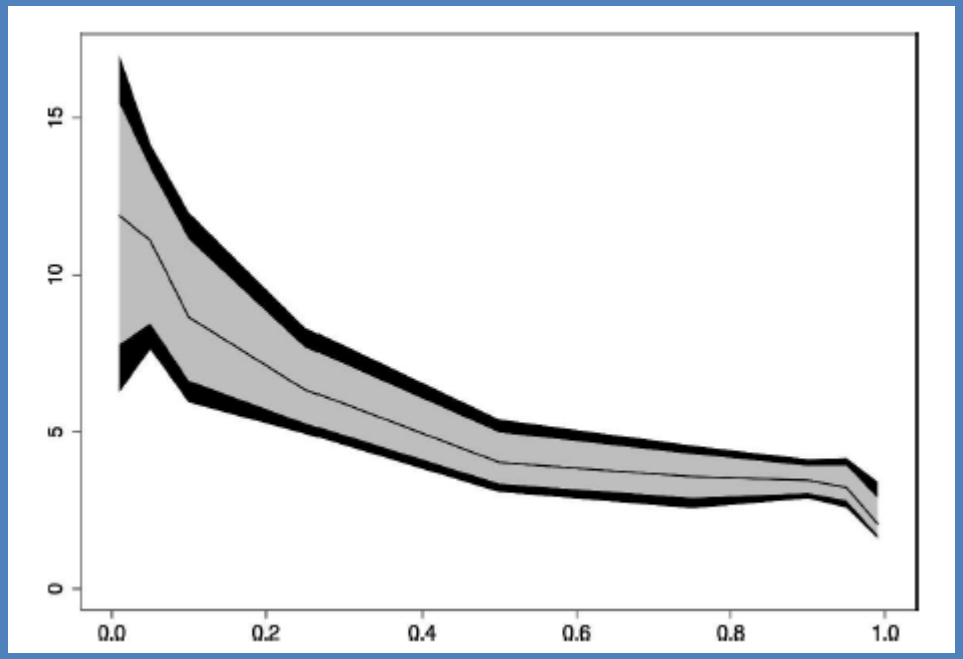
Ferro et al., 2005. IJoC

Confidence intervals

The interval $[L_p, U_p]$ is a pointwise confidence interval for d_p if:

$$P(L_p \leq d_p \leq U_p) = 1 - \alpha$$

$$d_p = y_p - x_p$$



L_p and U_p can be defined based on bootstrap resampling

Pointwise versus simultaneous confidence intervals

MEASURES

The measures to validate the occurrence of events are the hit rate and the false alarm rate,which are summarised in contingency tables (e.g. Wilks 1995, Jolliffe & Stevenson, 2003).

Continuous variables can be compared using these measures by defining suitable thresholds.

Table 3.1 The four possible outcomes for categorical forecasts of a binary event

Event forecast	Event observed	
	Yes	No
Yes	Hit	False alarm
No	Miss	Correct rejection

FINLEY AFFAIR

Table 3.1 The four possible outcomes for categorical forecasts of a binary event

Event forecast	Event observed	
	Yes	No
Yes	Hit	False alarm
No	Miss	Correct rejection

Table 1.1 Finley's Tornado forecasts

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

Finley (1884) assessed his performance using percent correct measure:

(Hits + correct rejections)/ All = 96.6%

Gilbert (1884) hasshown that he could get better performance giving always forecast „no tornado”:

No tornado/All = 98.2%

MEASURES

Several downscaling approaches predict local-scale probability density distributions rather than specific values.

Their performance can be validated by probability scores.

The classic measure to validate the occurrence of events is the Brier score (Brier 1950).

$$B = \frac{1}{n}\sum_{j=1}^{n}\left(p_j - x_j\right)^2$$

MEASURES

Continuous events (i.e. intensities) can be validated by the continuous ranked probability score (e.g. Jolliffe and Stephenson 2003) and the quantile verification score (e.g. Friederichs and Hense 2007).

$$RPS = E\left[\frac{1}{K}\sum_{k=1}^{K}\left(p_k - o_k\right)^2\right] = \frac{1}{K}\sum_{k=1}^{K} B_k$$

K > 2 number of thresholds $x_1 < x_2 < \ldots. < x_K$, defining events $A_k = \{X < x_k\}$, $p_1, p_2, \ldots, p_K$, forecast probabilities of events $A_1, A_2, \ldots, A_K$. o_k=1, if A_k occurs and 0 otherwise. B_k is a Brier score for event A_K.

$$CRPS = E\left(\int_{-\infty}^{\infty}\left[F(x) - H(x - x_0)\right]^2 dx\right)$$

F = c.d.f.
H = Heaviside function

THANKS