Ensemble Forecasts and their Verification

Expanded material to Sects. 6.5 and 7.4 of Wilks

Malaquias Peña
IMSG at EMC/NCEP/NOAA

Acknowledgments:
Eugenia Kalnay, notes on verification of EPS from Zoltan Toth and Yuejian Zhu (NOAA), Renate Hagedorn (ECMWF), and Laurence Wilson (EC). Ensemble Group at EMC.

AOSC630 Guest Class

March 30st, 2011

Outline

- Background
  - Uncertainty in NWP systems, Ensembles
- Ensemble forecast verification
  - Performance metrics: Brier Score, CRPSS
- New concepts and developments
  - Distribution fitters, Multi-model combination

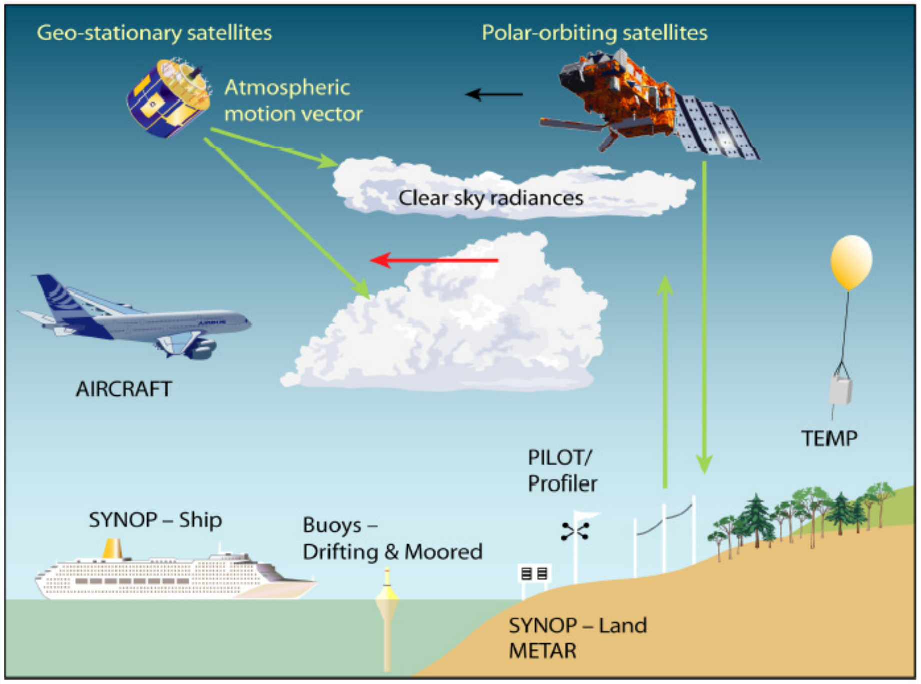
Background on Ensemble Forecasting

- Uncertainty in NWP systems
- Ensemble forecasts and probability forecasts
- Ensemble products
  - Ensemble mean
  - Spaghetti Plots
  - Stamps and plumes

Uncertainties in NWP

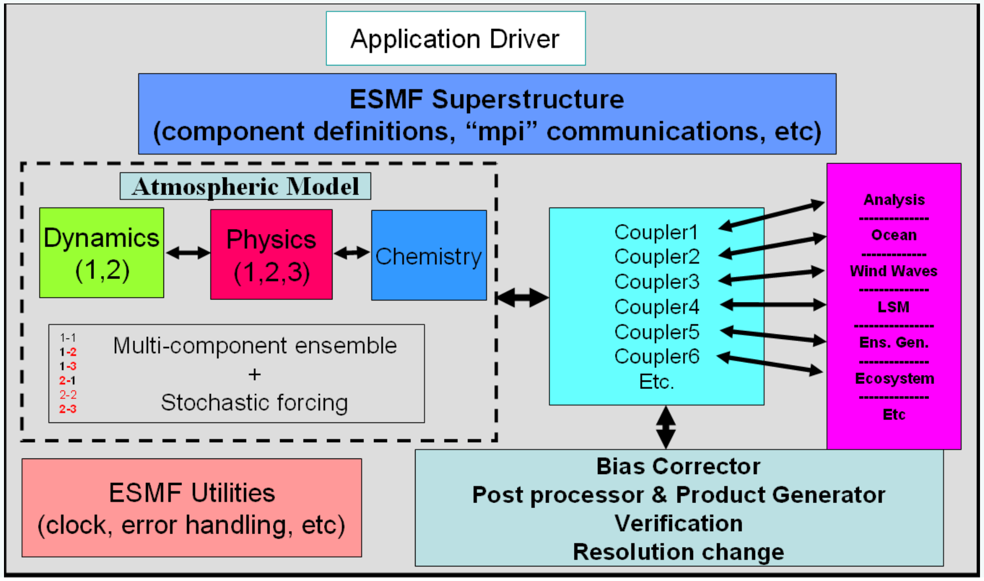
Observation errors → Uncertainty in initial conditions

Sources: Instrument errors, bias in frequency of measurements, representativeness, reporting errors, random errors, precision errors, conversion errors, etc.


Geo-stationary satellites
Polar-orbiting satellites
Atmospheric motion vector
Clear sky radiances
AIRCRAFT
TEMP
PILOT/ Profiler
SYNOP – Ship
Buoys – Drifting & Moored
SYNOP – Land METAR

Model errors → Uncertainty in modeling the evolution of weather

Sources: Insufficient spatial resolution, truncation errors in the dynamical equations, approximation errors to solve them, ad hoc parameterization, average errors, coding errors!, bias in frequency of initialization, etc.



Application Driver

ESMF Superstructure (component definitions, “mpi” communications, etc)

Atmospheric Model

Dynamics (1,2)

Physics (1,2,3)

Chemistry

1-1 1-2 1-3 2-1 2-2 2-3 Multi-component ensemble + Stochastic forcing

Coupler1 Coupler2 Coupler3 Coupler4 Coupler5 Coupler6 Etc.

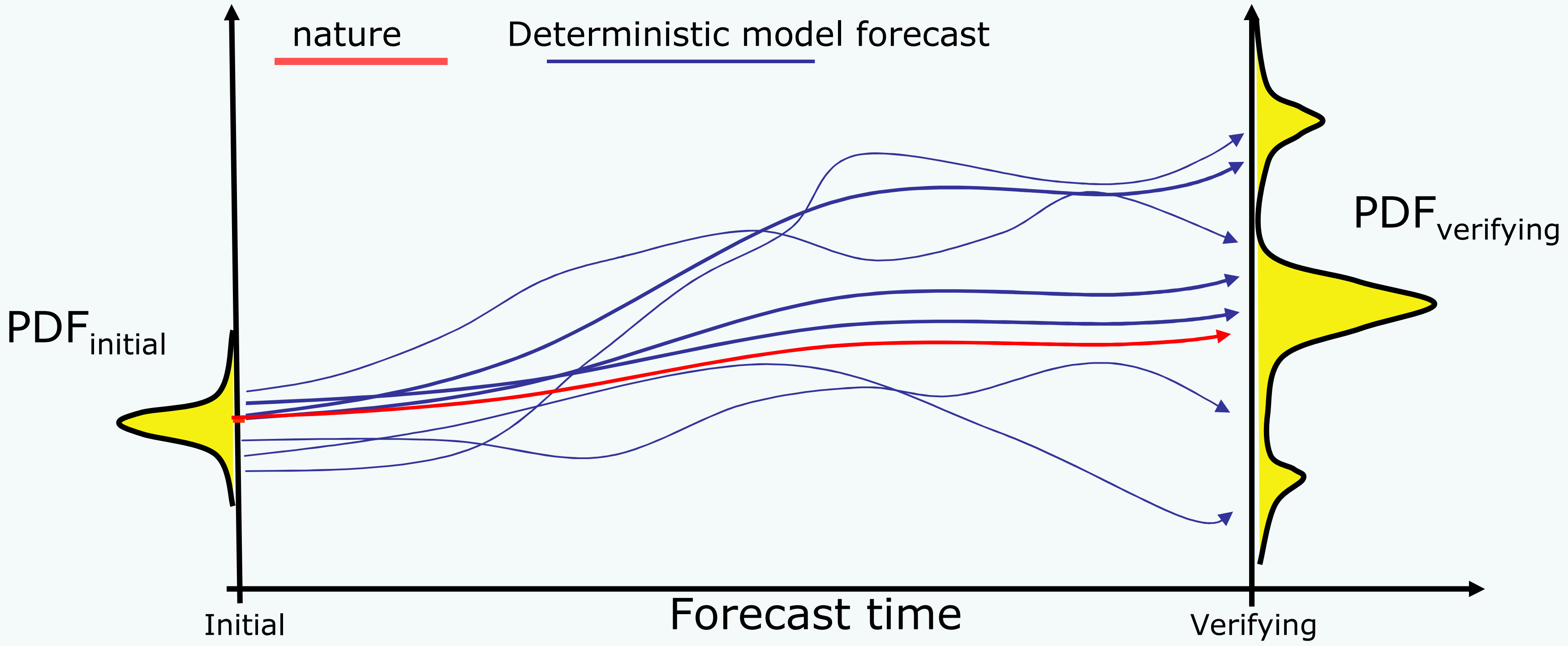
Analysis
Ocean
Wind Waves
LSM
Ens. Gen.
Ecosystem
Etc

ESMF Utilities (clock, error handling, etc)

Bias Corrector
Post processor & Product Generator
Verification
Resolution change

What is an ensemble forecast?

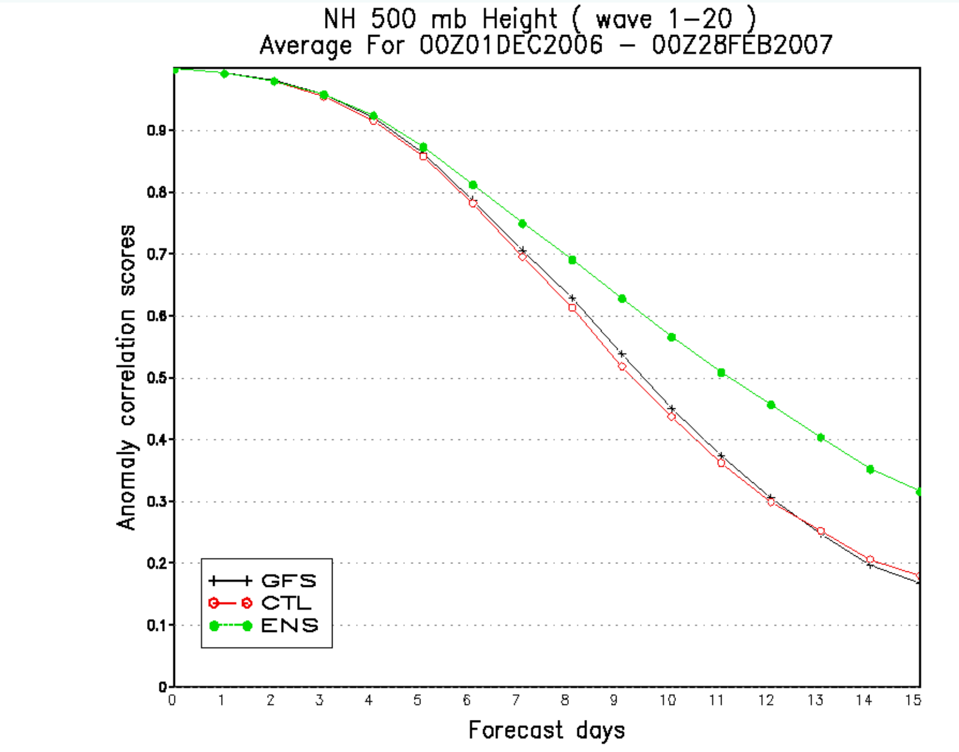
An approximation of the probability distribution reflecting the uncertainty associated with initial and model errors.



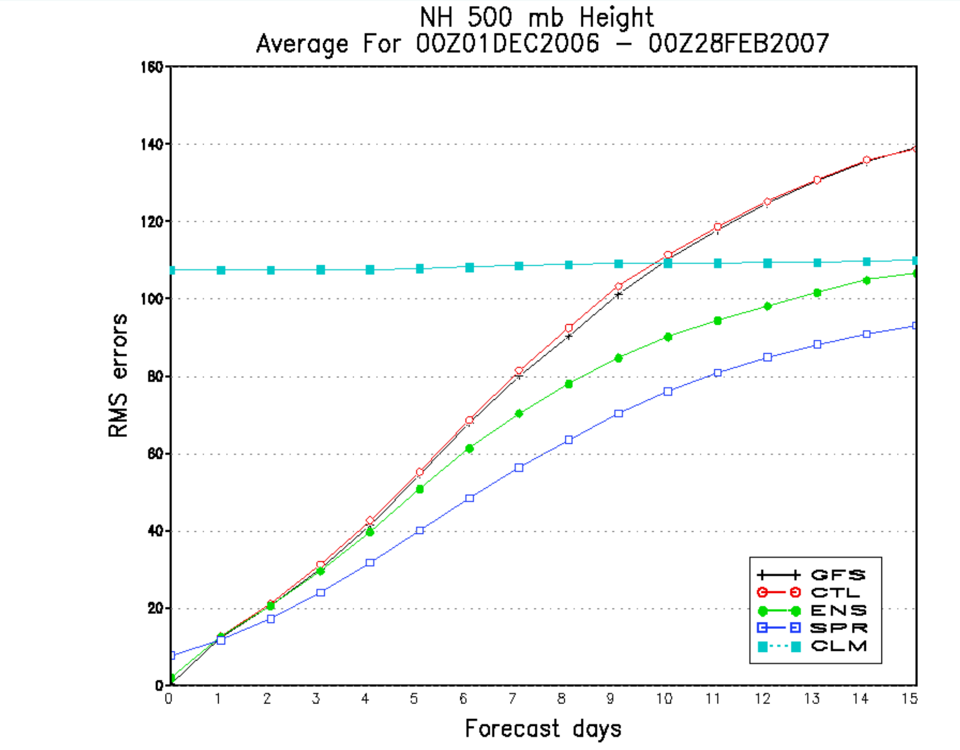
Solutions of NWP systems are sensitive to initial conditions; each model run samples the initial unknown PDF; an evolving uncertainty PDF is determined by collecting forecasts verifying at a given (verifying) time

Application: Ensemble Average

- Ensemble mean (=average when each ensemble member is equally likely to occur) can be used as single forecast.
- Average removes short-lived variations retaining slowly-varying patterns (Leith 1974).



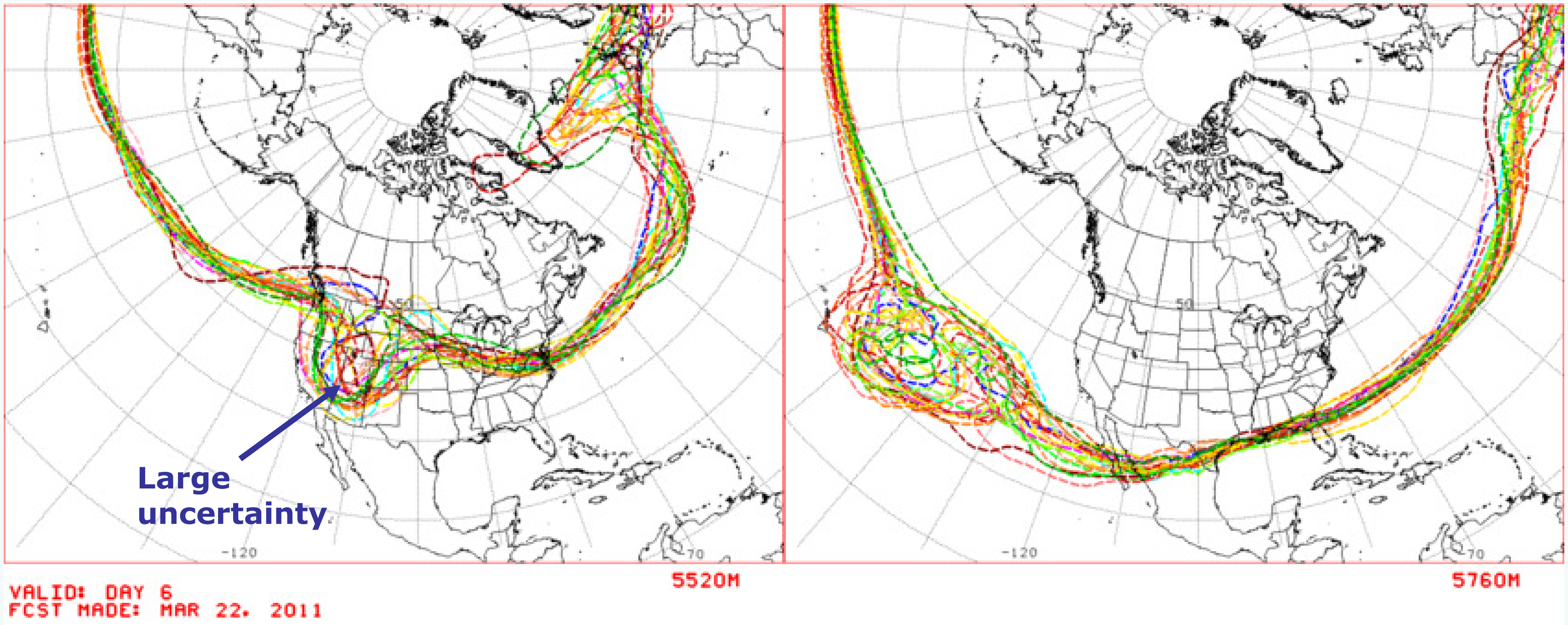
Ensemble mean (green) more skillful than single forecasts after day 4



After lead 4 errors from ensemble mean (green) approach climatology and are smaller than errors from single forecasts.

Figs. from Yuejian Zhu (EMC/NCEP/NOAA)

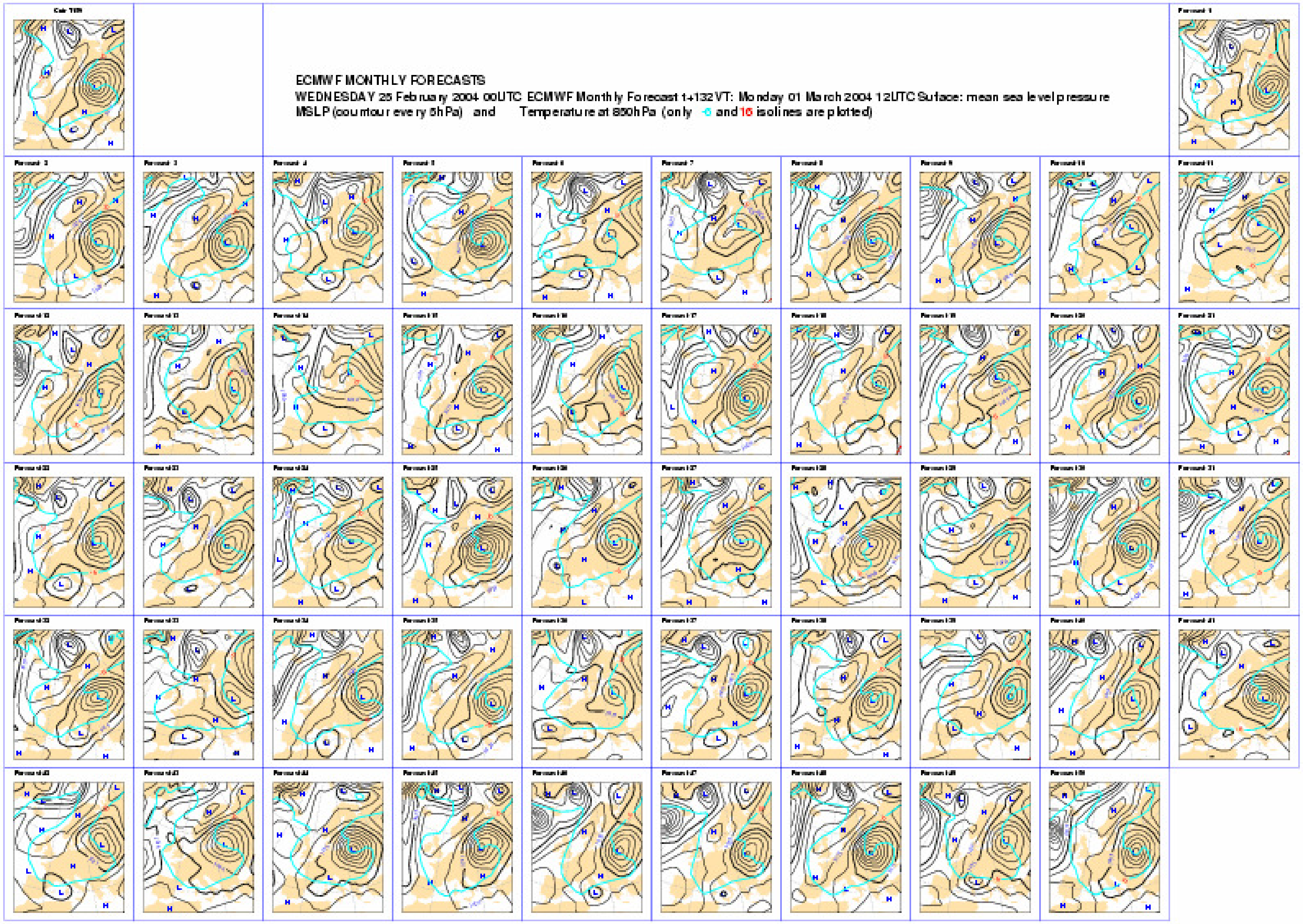
Spaghetti Diagrams



Large uncertainty

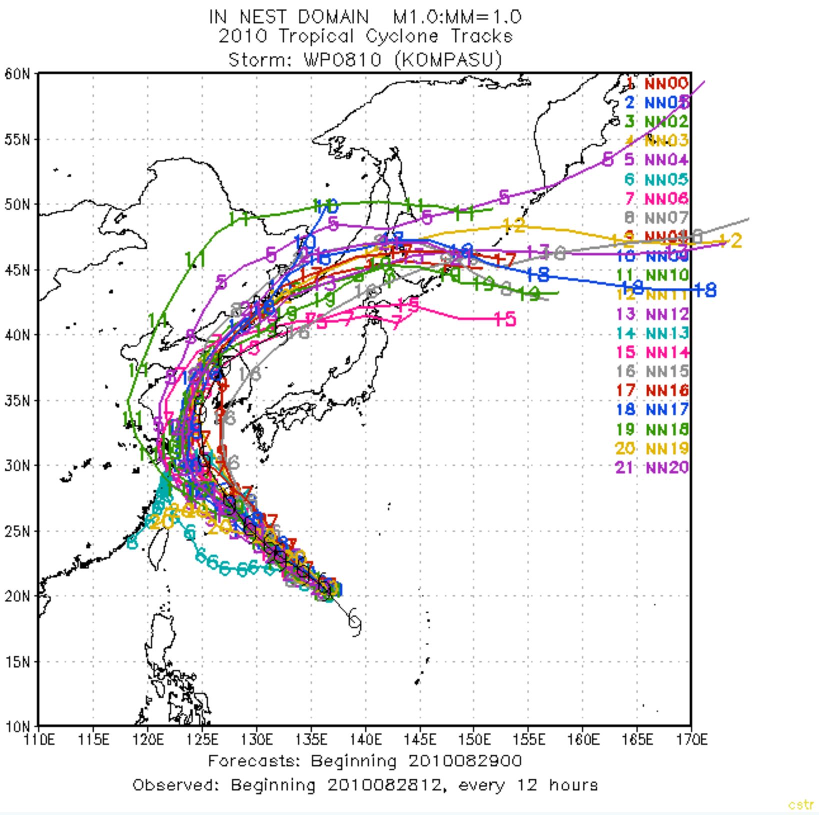
- Contours of ensemble forecasts at specific geopotential heights at 500hPa
- Visualizing the amount of uncertainty among ensemble members
    - High confidence of the forecast in regions where members tend to coincide
- Advantages over mean-spread diagrams: keeps features sharp, allows identifying clustering of contours (e.g., bi-modal distributions)

Stamps

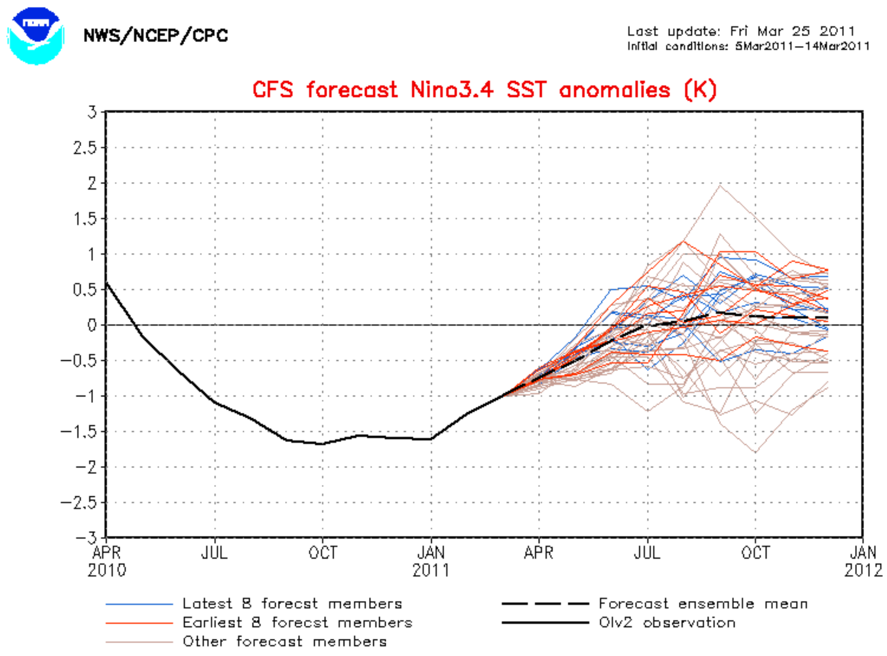


ECMWF MONTHLY FORECASTS
WEDNESDAY 25 February 2004 00UTC ECMWF Monthly Forecast t+132 VT: Monday 01 March 2004 12UTC Suface: mean sea level pressure
MSLP (countour every 5hPa) and Temperature at 850hPa (only 0 and 16 isolines are plotted)

Plumes



Hurricane tracks



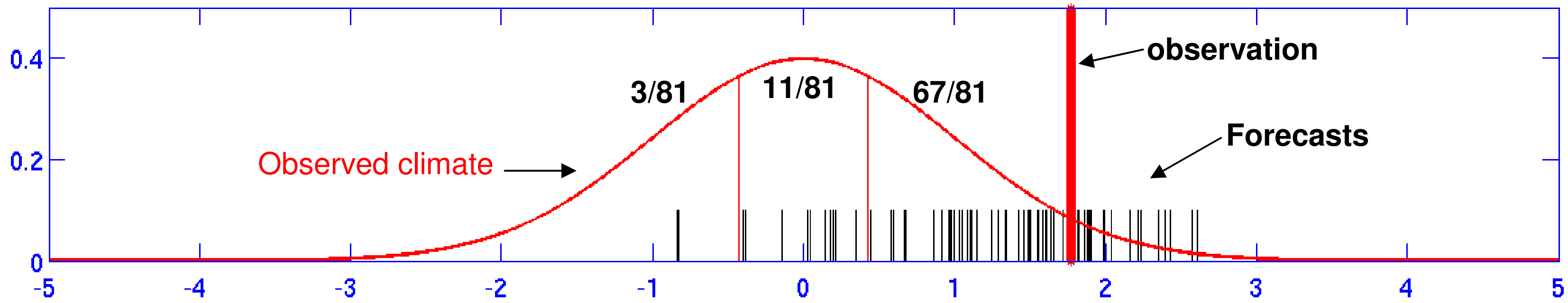
El Nino 3.4 Index

Quantitative description

- Assume each deterministic forecast in the ensemble is an independent realizations of the same random process
- Forecast probability of an event is estimated as the fraction of the forecasts predicting the event among all forecasts considered (relative frequency of occurrence)

If n_t is the ensemble size, and n_x is the number of members that predict the event x, the probability $P(x)$ that the event will happen can be approximated as:

$$P(x) \approx \frac{n_x}{n_t}$$



Probability forecast

- Probability conveys level of uncertainty of a given forecast

Categorical forecast: Yes/No. Only 100% or 0% probability	Probabilistic forecast: assigns a probability value between 0 and 100%

Example: There is a 30% chance of precipitation (referred to as PoP) for today in College Park

- Probabilistic forecasts cannot be verified with one single event
- Verification requires many cases in which the 30% PoP has been issued

Probabilistic forecast verification

- Comparison of a distribution of forecasts to a distribution of observations

- Characteristics of a forecast system:
  - **Reliability**: How well the *a priori* predicted probability forecast of an event coincides with the *a posteriori* observed frequency of the event
  - **Resolution**: How much the forecasts differ from the climatological mean probabilities of the event, and the systems gets it right?
  - **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
  - **Skill**: How much better are the forecasts compared to a reference prediction system (chance, climatology, persistence,…)?

Performance measures of probabilistic forecast

Brier Skill Score (BSS)
Reliability Diagrams
Relative Operating Characteristics (ROC)
Rank Probability Score (RPS)
Continuous RPS (CRPS)
CRP Skill Score (CRPSS)
Rank histogram (Talagrand diagram)

The Brier Score

- Mean square error of a probability forecast

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

where N is the number of realizations, *p_i* is the probability forecast of realization *i*. *O_i* is equal to 1 or 0 depending on whether the event (of realization *i*) occurred or not.

- Measures accuracy. Range: 0 to 1. Perfect=0
- Weights larger errors more than smaller ones

Components of the Brier Score

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

Decomposed into 3 terms for *K* probability classes and a sample of size *N*:

$$BS = \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \overline{o}_k)^2}_{\text{reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k (\overline{o}_k - \overline{o})^2}_{\text{resolution}} + \underbrace{\overline{o}(1-\overline{o})}_{\text{uncertainty}}$$

reliability

If for all occasions when forecast probability p_k is predicted, the observed frequency of the event is

$$\overline{o}_k = p_k$$

then the forecast is said to be reliable. Similar to bias for a continuous variable

resolution

The ability of the forecast to distinguish situations with distinctly different frequencies of occurrence.

uncertainty

The variability of the observations. Maximized when the climatological frequency (*base rate*) =0.5

Has nothing to do with forecast quality! Use the Brier skill score to overcome this problem.

The presence of the uncertainty term means that Brier Scores should not be compared on different samples.

Brier Skill Score

Skill: Proportion of improvement of accuracy over the accuracy of a reference forecast (e.g., climatology or persistence)

- Brier Skill Score

$$BSS = -\frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{ref}}}$$

- If the sample climatology is used, BSS can be expressed as:

$$BSS = -\frac{Res - Rel_{ref}}{Unc_{ref}}$$

- Range: -Inf to 1; No skill beyond reference=0; Perfect score =1

Brier Score and Skill Score - Summary

- Measures accuracy and skill respectively
- Cautions:
  - Cannot compare BS on different samples
  - BSS – Take care about underlying climatology
  - BSS – Take care about small samples

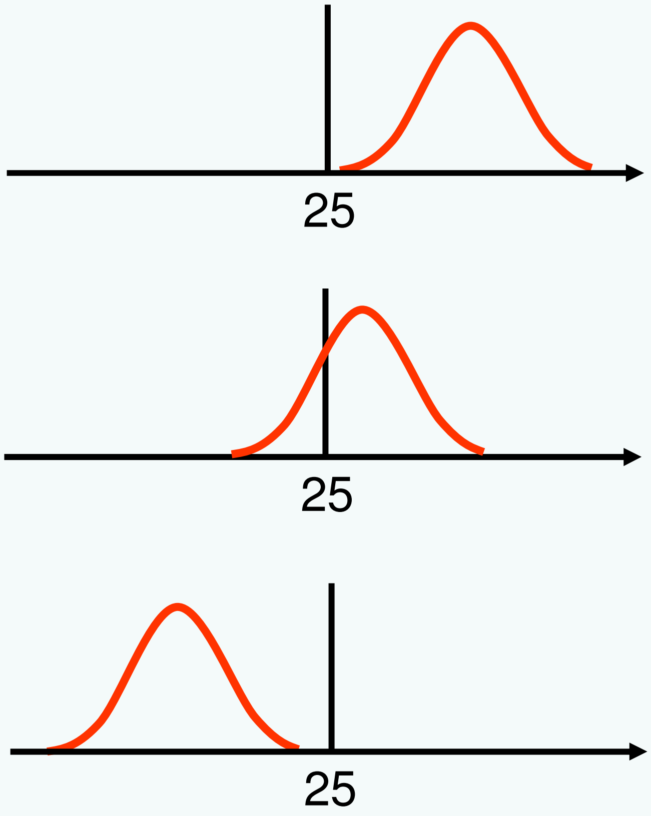
Reliability

- A forecast system is reliable if:
  - statistically the predicted probabilities agree with the observed frequencies, i.e. taking all cases in which the event is predicted to occur with a probability of x%, that event should occur exactly in x% of these cases; not more and not less.
  - Example: Climatological forecast is reliable but does not provide any forecast information beyond climatology

- A reliability diagram displays whether a forecast system is reliable (unbiased) or produces over-confident / under-confident probability forecasts
- A reliability diagram also gives information on the resolution (and sharpness) of a forecast system

Reliability Diagram

Take a sample of probabilistic forecasts:
e.g. 30 days x 2200 GP = 66000 forecasts
How often was event (T > 25) forecasted with X probability?

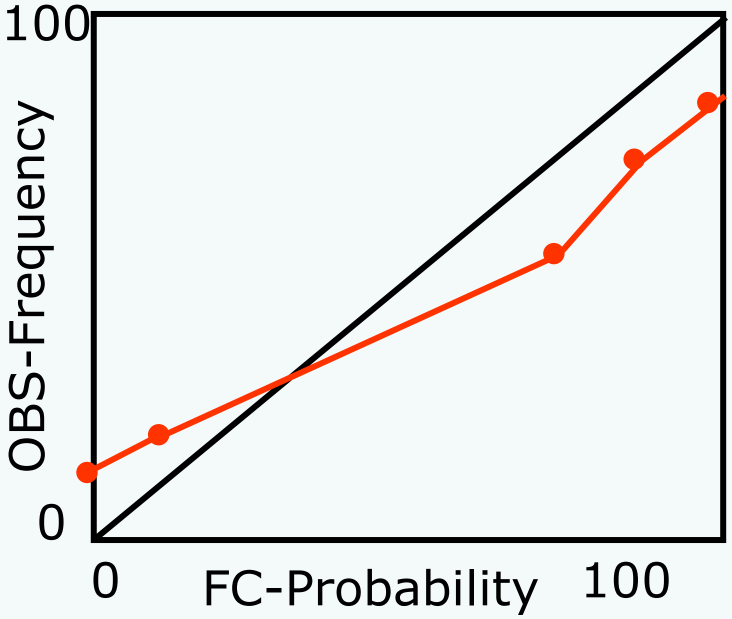


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
....	....	....	....
....	....	....	....
....	....	....	....
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

R. Hagedorn, 2007

Reliability Diagram

Take a sample of probabilistic forecasts:
e.g. 30 days x 2200 GP = 66000 forecasts
How often was event (T > 25) forecasted with X probability?

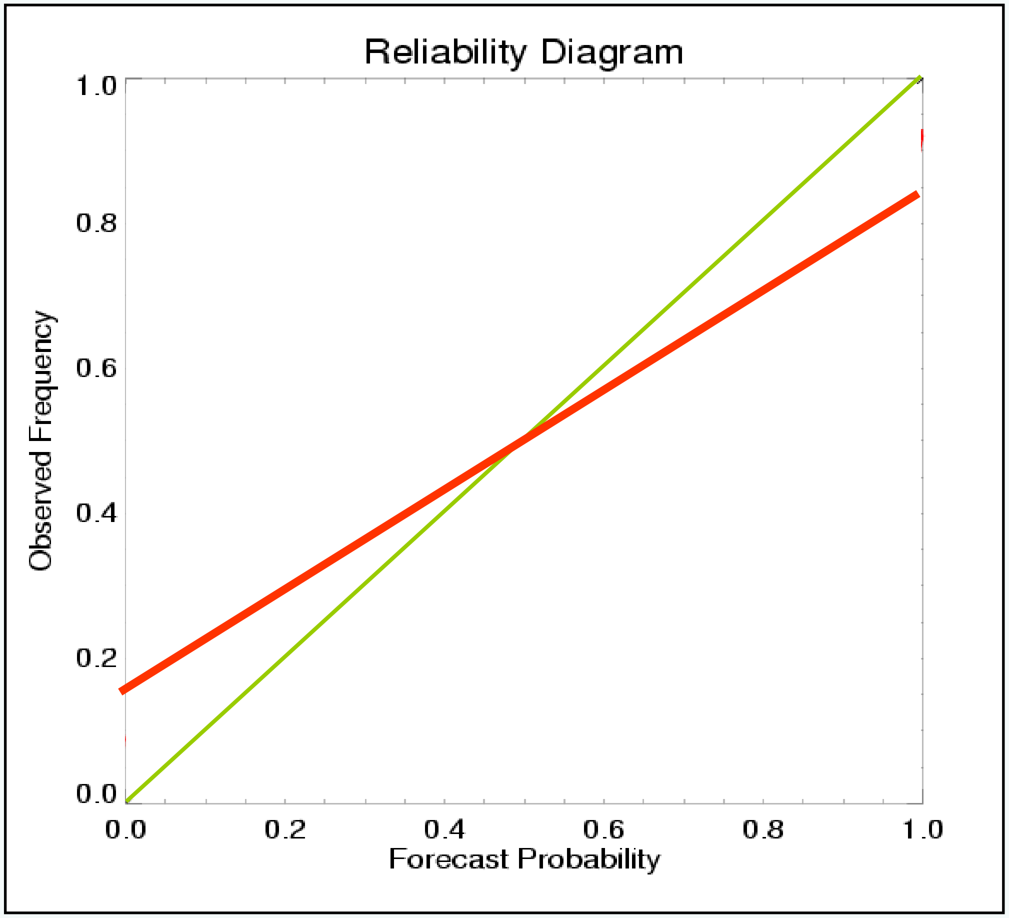


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

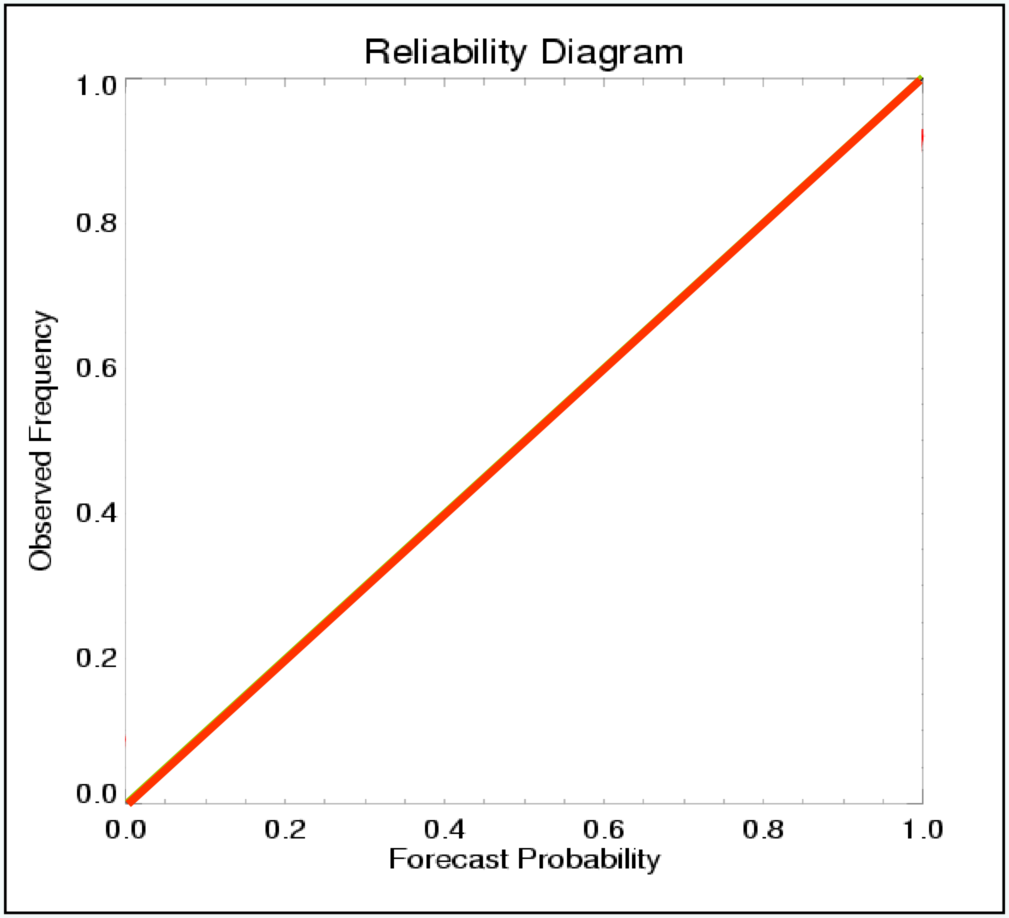
R. Hagedorn, 2007

Reliability Diagram

over-confident model



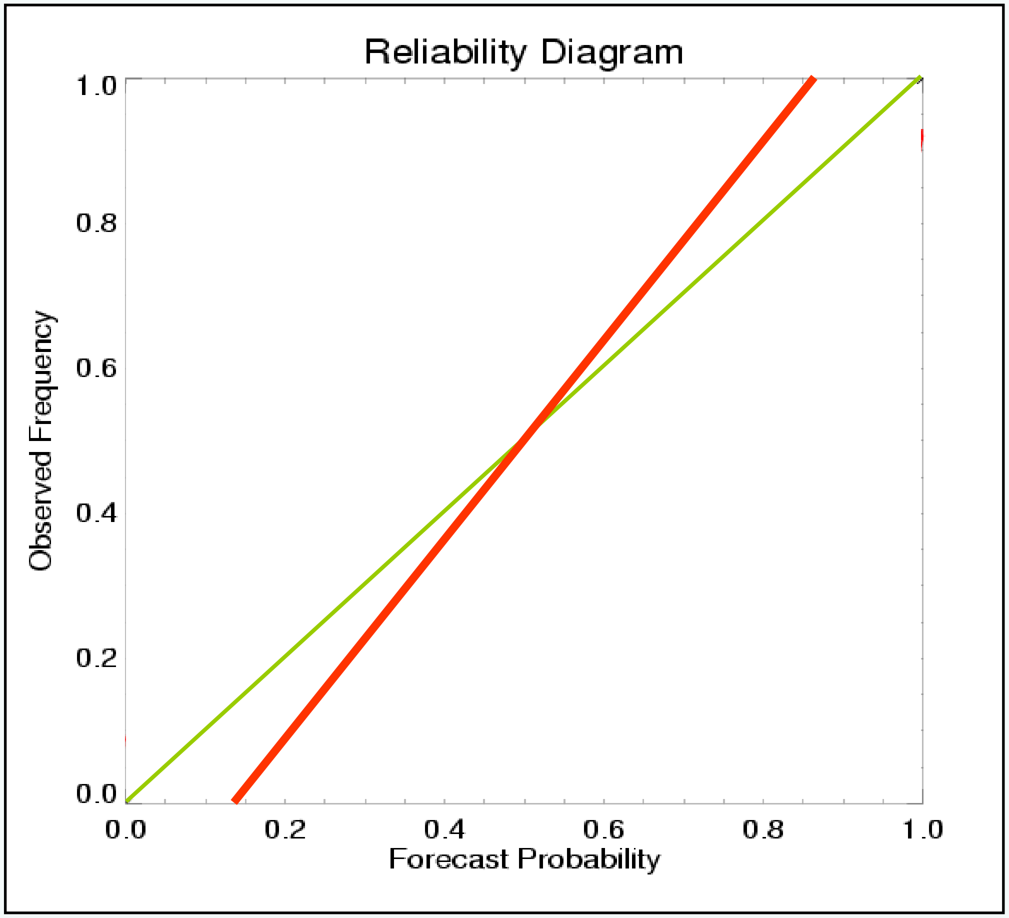
perfect model



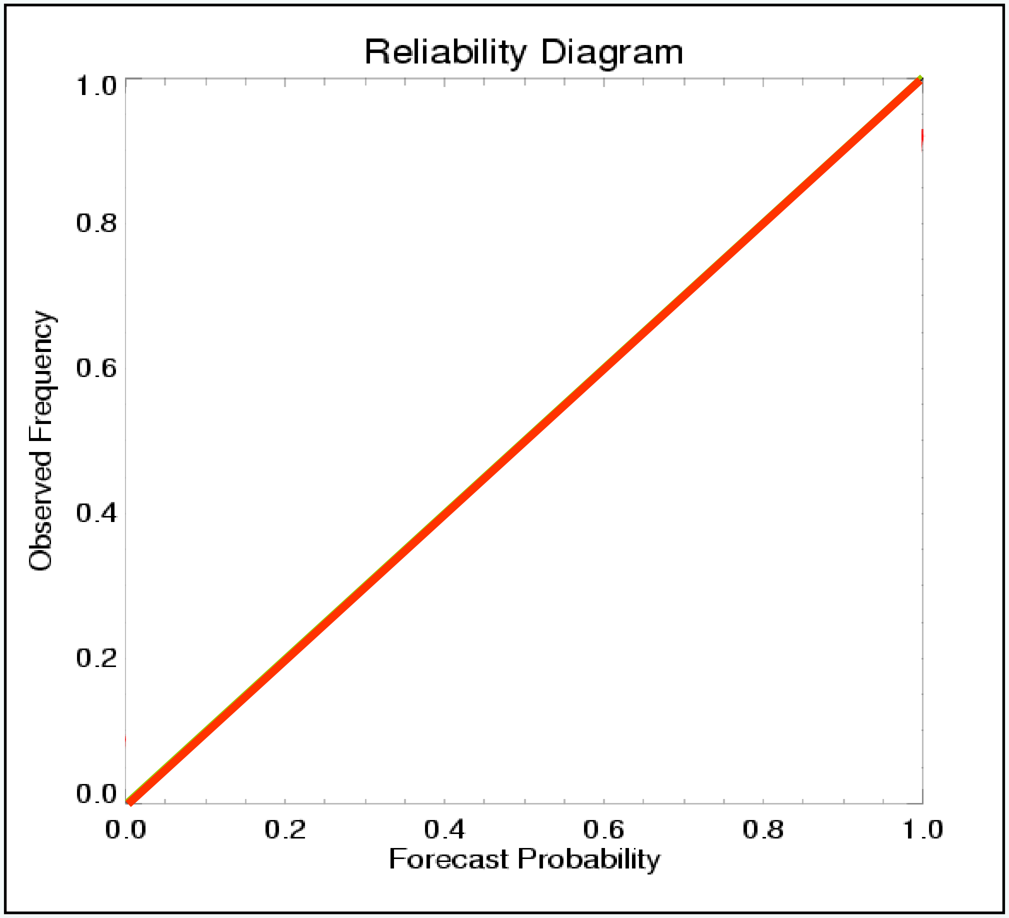
R. Hagedorn, 2007

Reliability Diagram

under-confident model



perfect model

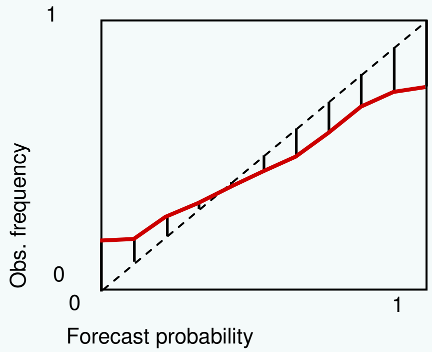


R. Hagedorn, 2007

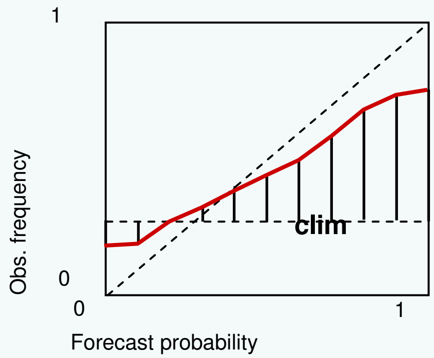
Reliability Diagram

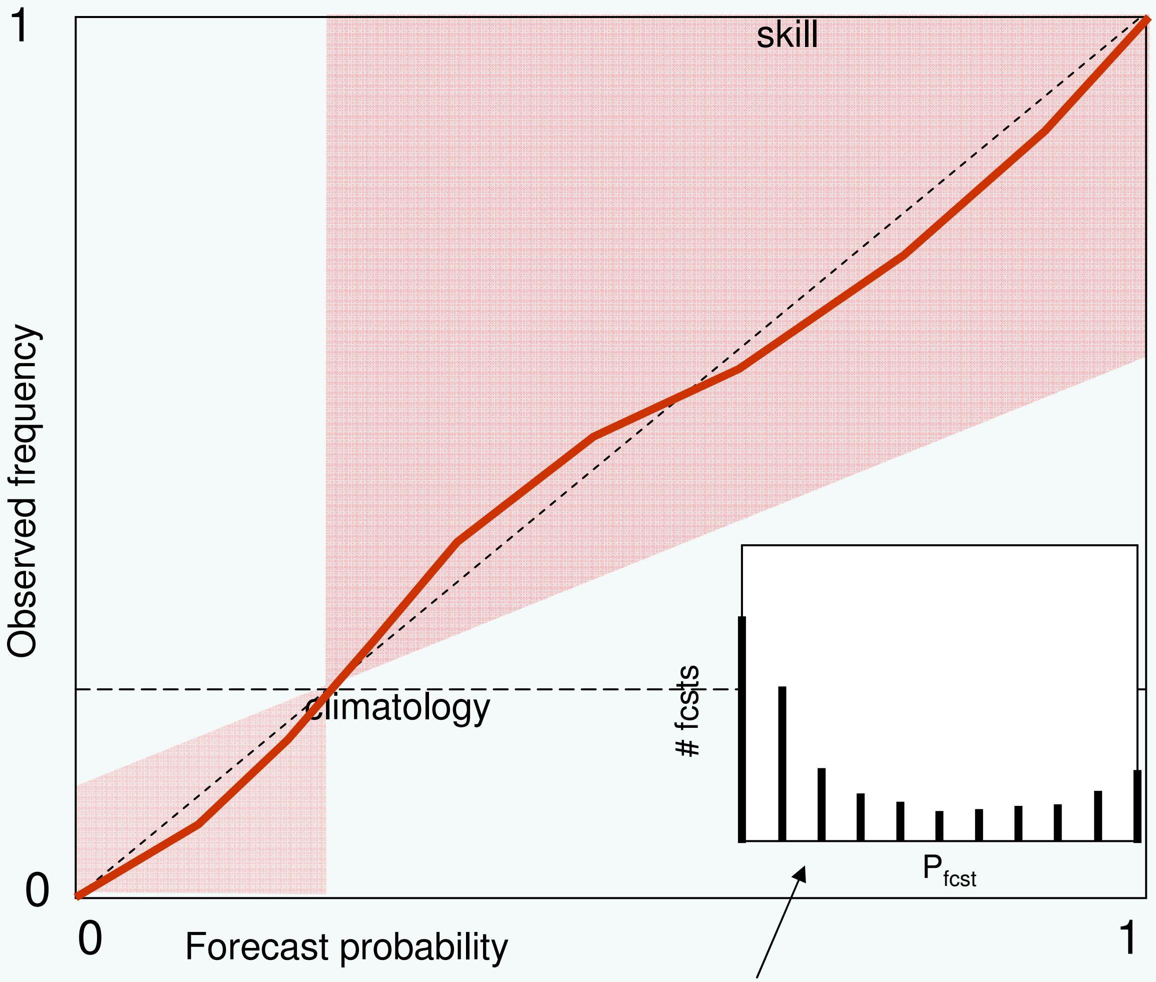
Reliability: Proximity to diagonal

(the lower the value the better)



Resolution: Proximity to horizontal (climatology) line





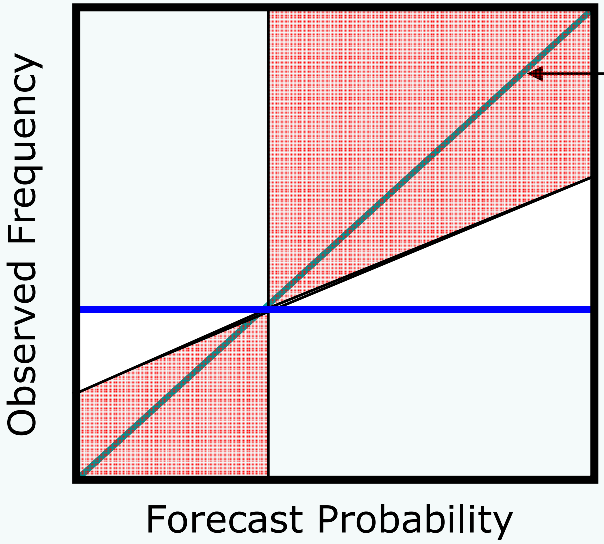
Sharpness diagram

Brier Skill Score & Reliability Diagram

- How to construct the area of positive skill?

$$BSS = 1 - \frac{BS}{BS_c}$$

$$= 1 - \frac{REL - RES + UNC}{UNC} = \frac{RES - REL}{UNC}$$



perfect reliability

line of no skill

climatological frequency (line of no resolution)

▨ area of skill (RES > REL)

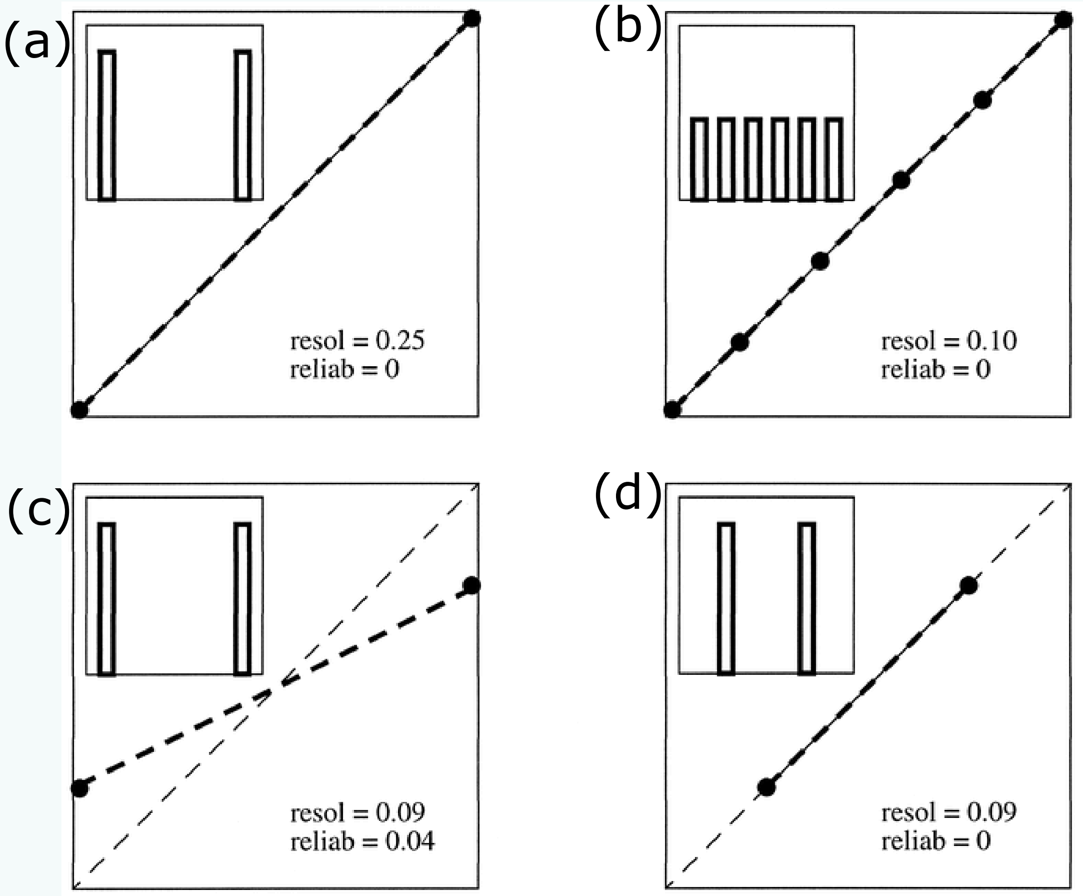
R. Hagedorn, 2007

Construction of Reliability diagram

1. Decide number of categories (bins) and their distribution:
   - Depends on sample size, discreteness of forecast probabilities
   - Should be an integer fraction of ensemble size
   - Don’t all have to be the same width – within bin sample should be large enough to get a stable estimate of the observed frequency.
2. Bin the data
3. Compute observed conditional frequency in each category (bin) *k*
   - *obs. relative frequency*$_k$ = *obs. occurrences*$_k$ / *num. forecasts*$_k$
4. Plot observed frequency vs forecast probability
5. Plot sample climatology ("no resolution" line) (The sample base rate)
   - *sample climatology* = *obs. occurrences* / *num. forecasts*
6. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
7. Plot forecast frequency histogram to show sharpness (or plot number of events next to each point on reliability graph)

Reliability Diagram Examples

Atger, 1999

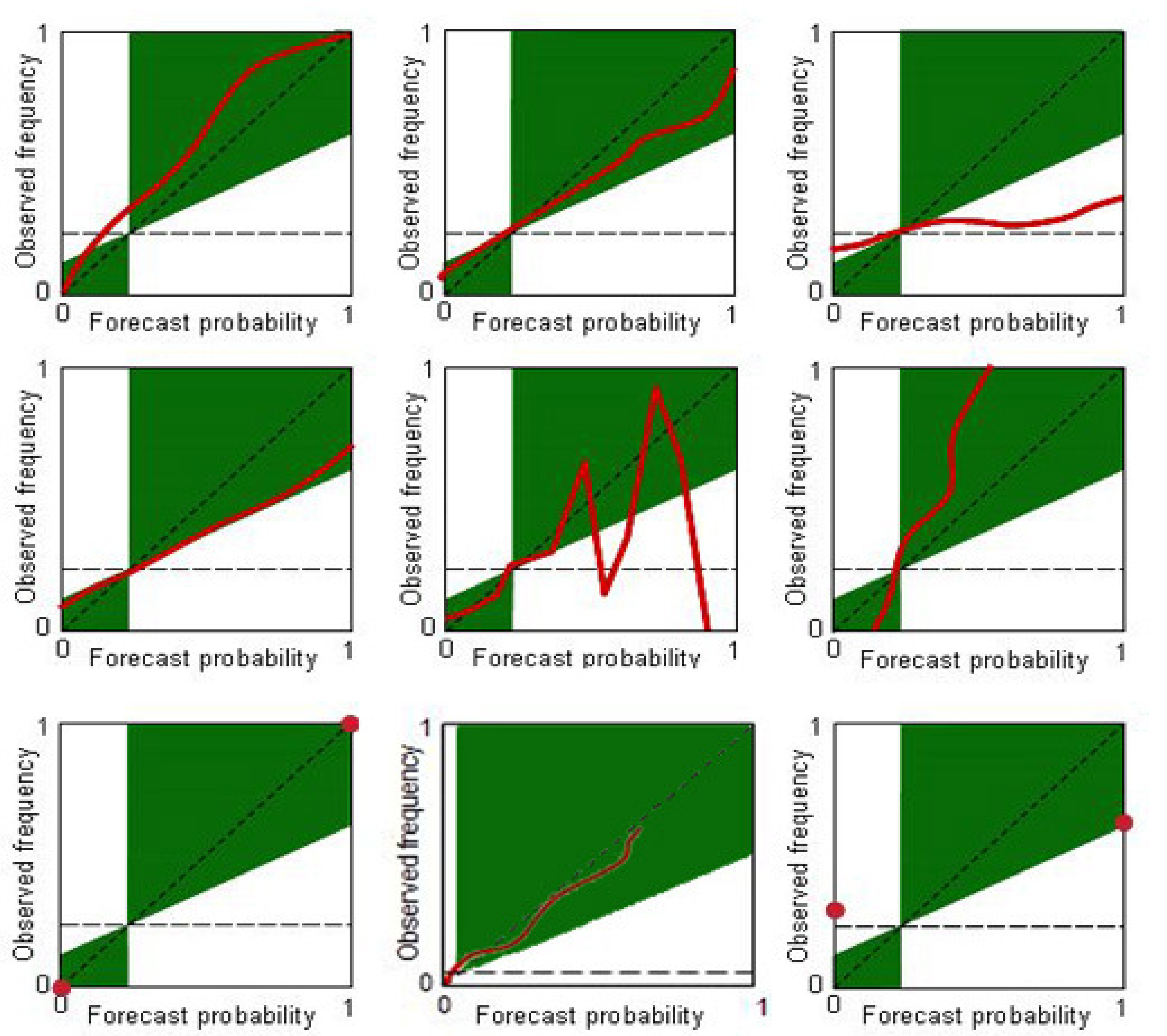


Typical reliability diagrams and sharpness histograms (showing the distribution of predicted probabilities). (a) Perfect resolution and reliability, perfect sharpness. (b) Perfect reliability but poor sharpness, lower resolution than (a). (c) Perfect sharpness but poor reliability, lower resolution than (a). (d) As in (c) but after calibration, perfect reliability, same resolution.

Reliability Diagram Exercise

Identify diagram(s) with:

1. Categorical forecast
2. Overconfident
3. Underconfident
4. Unskillful
5. Not sufficiently large sampling



From L. Wilson (EC)

Comments on Reliability Diagrams

- Requires a fairly large dataset, because of the need to partition (bin) the sample into subsamples conditional on forecast probability
- Sometimes called “attributes” diagram.

Reliability Diagrams - Summary

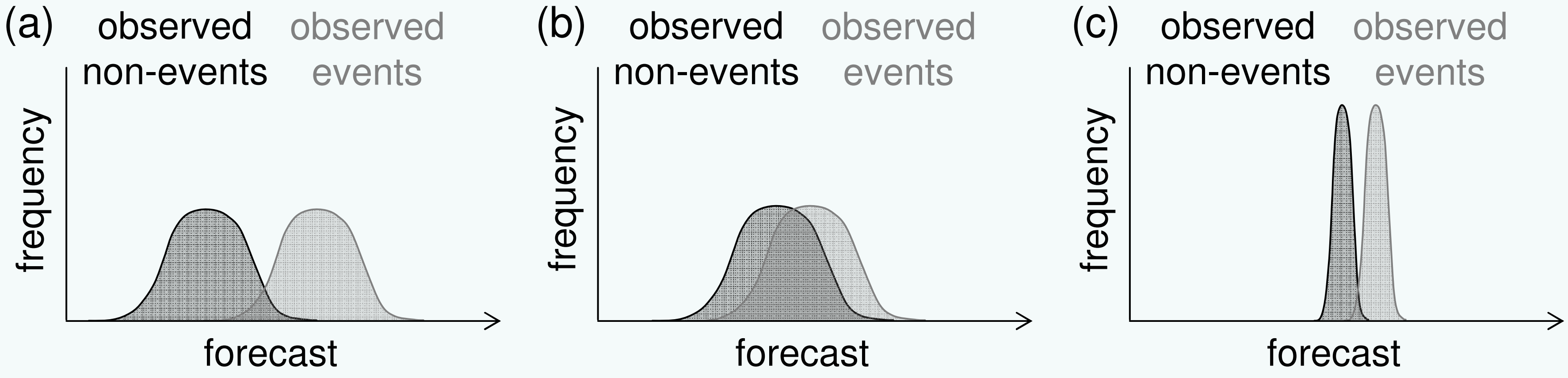
- Diagnostic tool
- Measures “reliability”, “resolution” and “sharpness”
- Requires “reasonably” large dataset to get useful results
- Try to ensure enough cases in each bin
- Graphical representation of Brier score components

Discrimination and the Relative Operating Curve

- Reliability diagram – partitioning the data according to the forecast probability
- Suppose we partition according to observation – 2 categories, yes or no
- Look at distribution of forecasts separately for these two categories

Discrimination

- *Discrimination*: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.
- Depends on:
  - Separation of means of conditional distributions
  - Variance within conditional distributions



Good discrimination

Poor discrimination

Good discrimination

From L. Wilson (EC)

Contingency Table

Observed

Forecast

	yes	no
yes	hits	false alarms
no	misses	Correct negatives

$$Accuracy = \frac{hits + correct\ negatives}{total}$$

False Alarm Rate

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

Hit Rate

$$HR = \frac{hits}{hits + misses}$$

$$BIAS = \frac{hits + false\ alarms}{hits + misses}$$

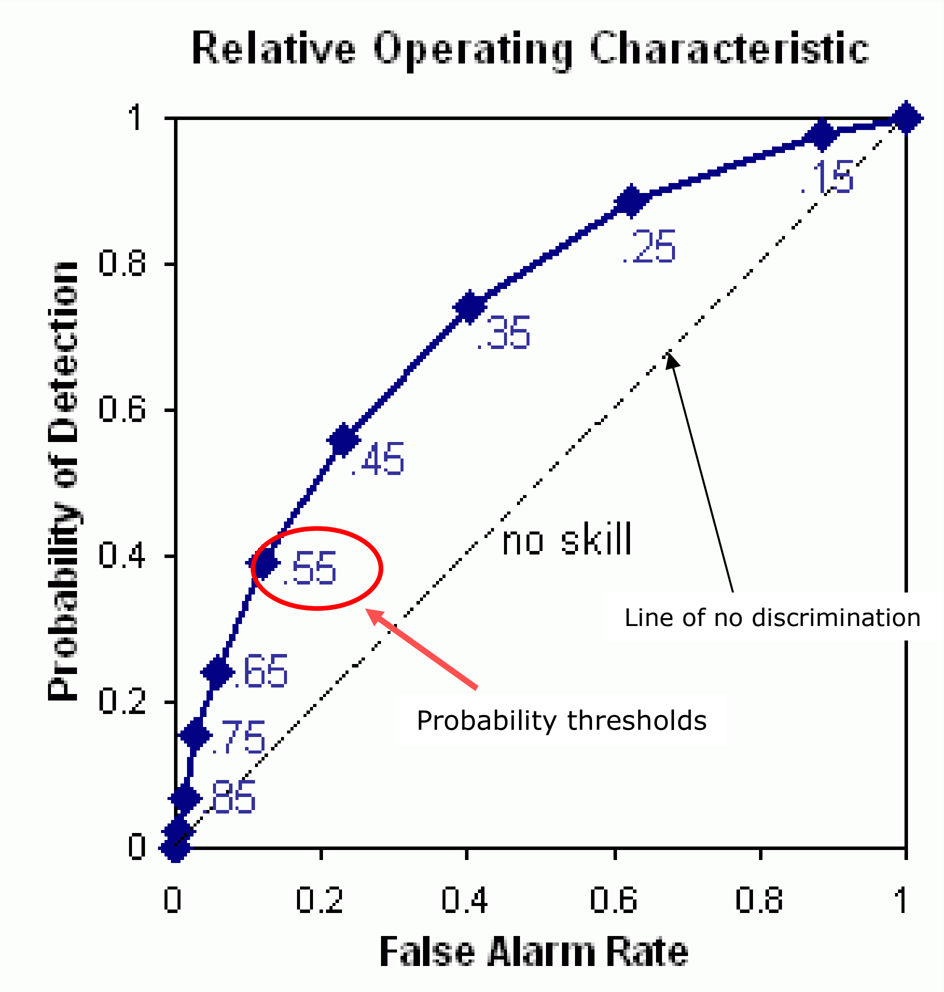
Construction of ROC curve

- From original dataset, determine bins
  - Can use binned data as for Reliability diagram BUT
  - There must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
  - Generally need at least 5 bins.
- For each probability threshold, determine HR and FAR
- Plot HR vs FAR to give empirical ROC.
- Obtain ROC area

ROC - Interpretation

- Measures ability to discriminate between two possible outcomes
- Measures resolution; it says nothing about reliability (ROC is not sensitive to bias)
- Area under ROC curve (A) used as a single quantitative measure
- Area range: 0 to 1. Perfect =1. No Skill = 0.5
- ROC Skill Score (ROCSS)

$$\text{ROCSS} = 2A - 1$$



Comments on ROC

- Measures “discrimination”
- The ROC is conditioned on the observations (i.e., given that Y occurred, what was the corresponding forecast?) It is therefore a good companion to the reliability diagram, which is conditioned on the forecasts.
- Sensitive to sample climatology – careful about averaging over areas or time
- Related to the assessment of “value” of forecasts
- Can compare directly the performance of probability and deterministic forecast

Rank Probability Score

- Measures the quadratic distance between forecast and verification probabilities for several probability categories *k*. Range: 0 to 1. Perfect=0

- Emphasizes accuracy by penalizing large errors more than “near misses”

- Rewards sharp forecast if it is accurate

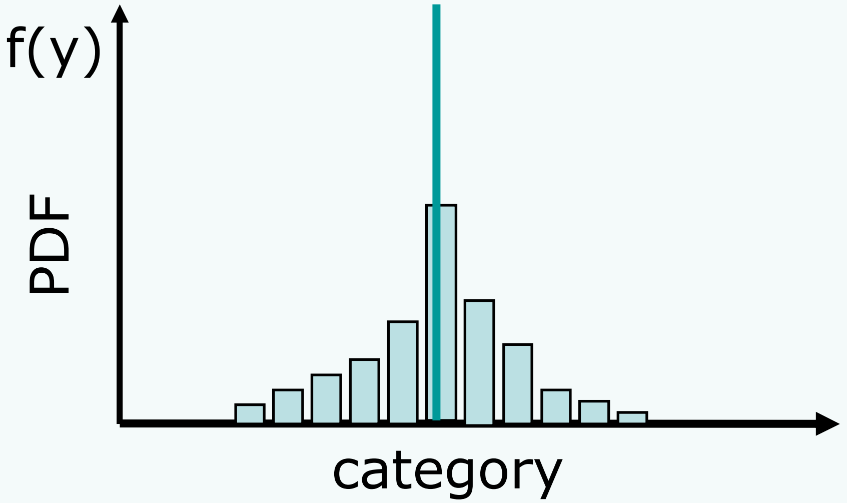
- It is the average Brier score across the range of the variable

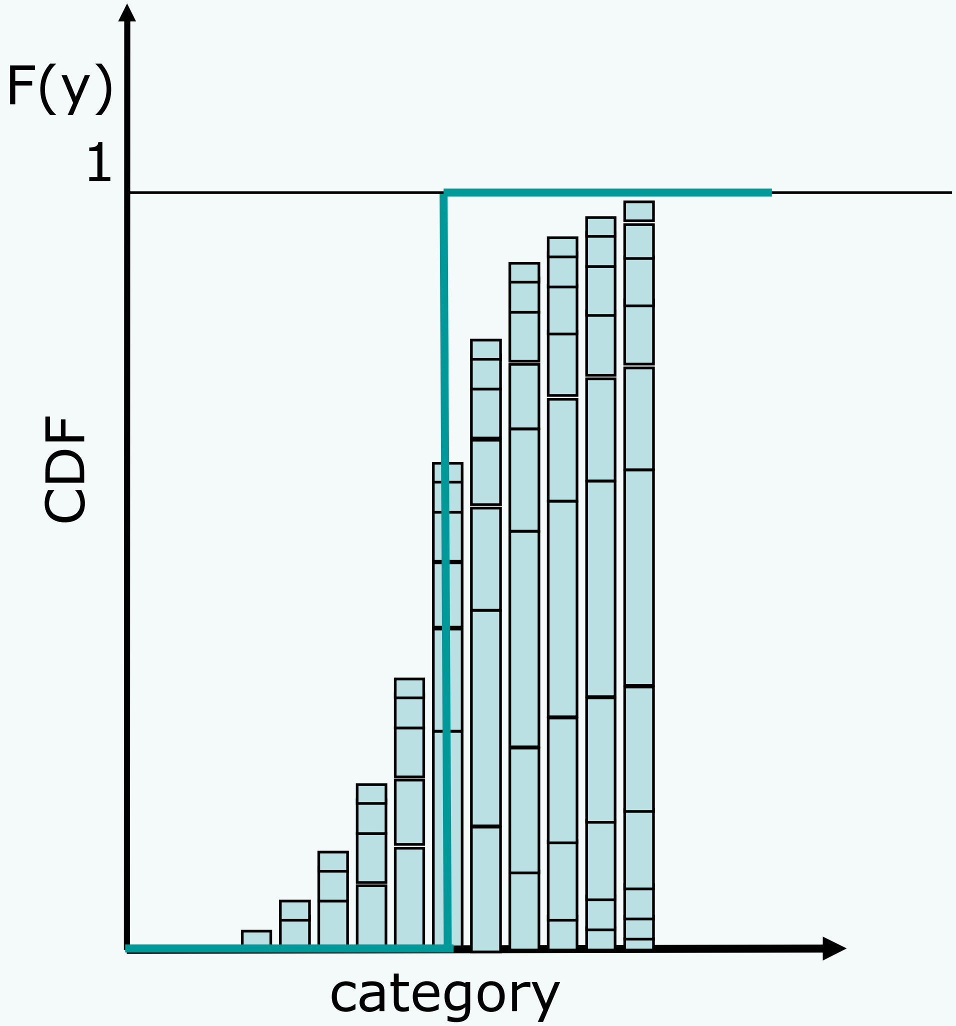
$$RPS = \frac{1}{K-1}\sum_{k=1}^{K} BS_k$$

- Ranked Probability Skill Score (RPSS) is a measure for skill relative to a reference forecast

$$RPSS = 1 - \frac{RPS}{RPS_{reference}}$$

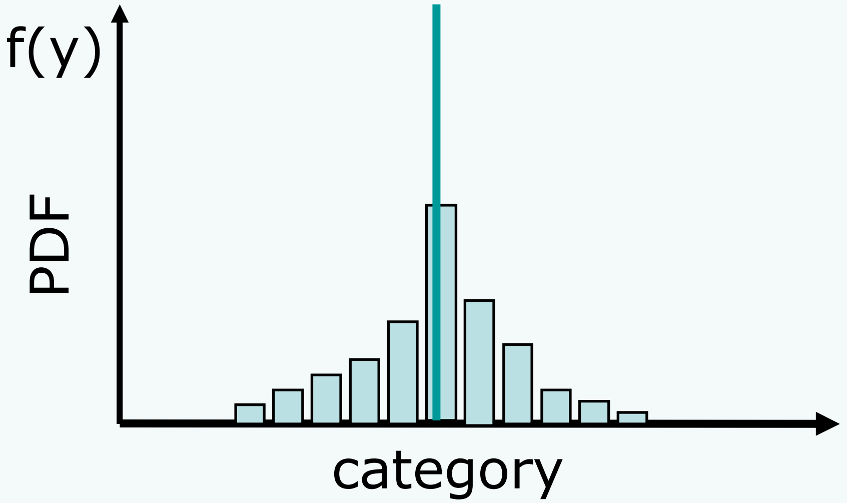
Rank Probability Score



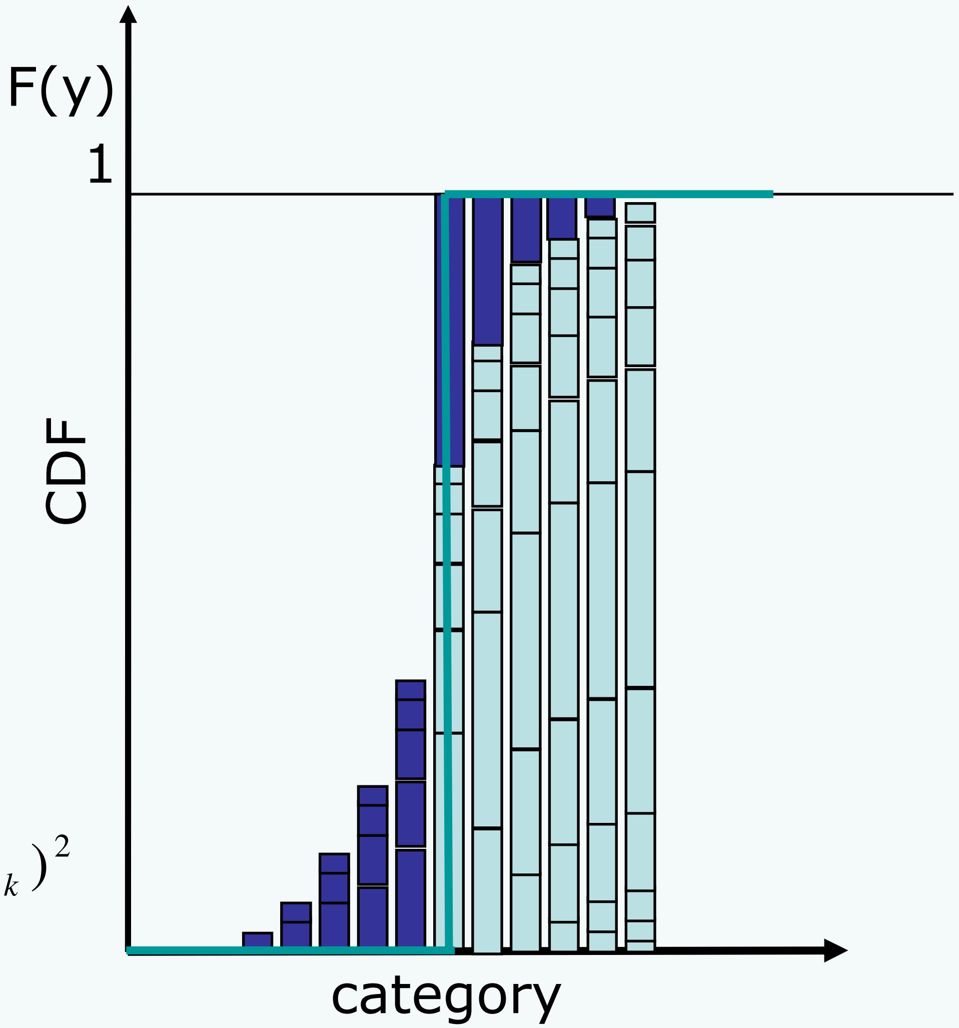


R. Hagedorn, 2007

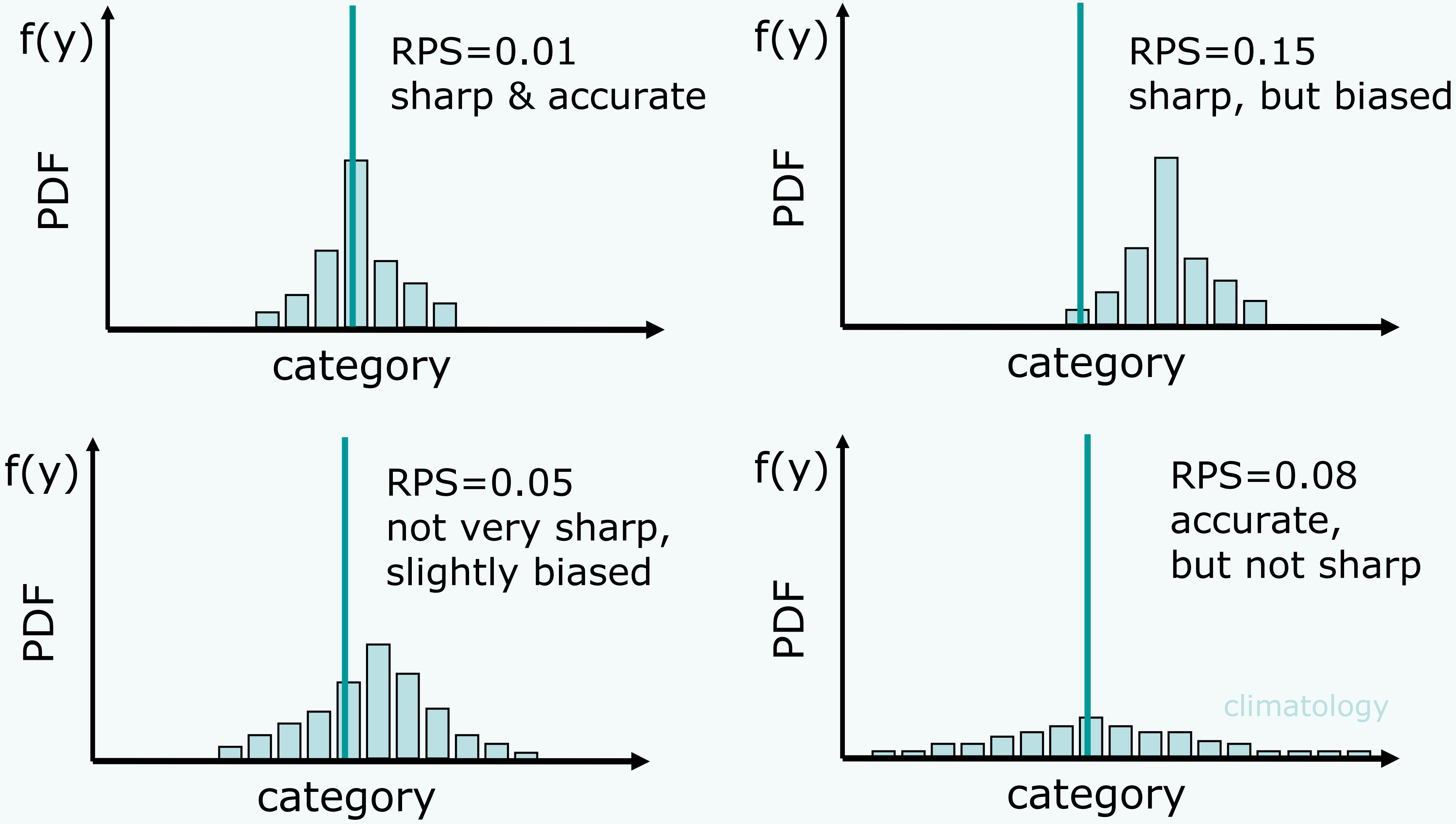
Rank Probability Score



$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{FC,k} - CDF_{OBS,k})^2$$



Rank Probability Score



RPS=0.01
sharp & accurate

RPS=0.15
sharp, but biased

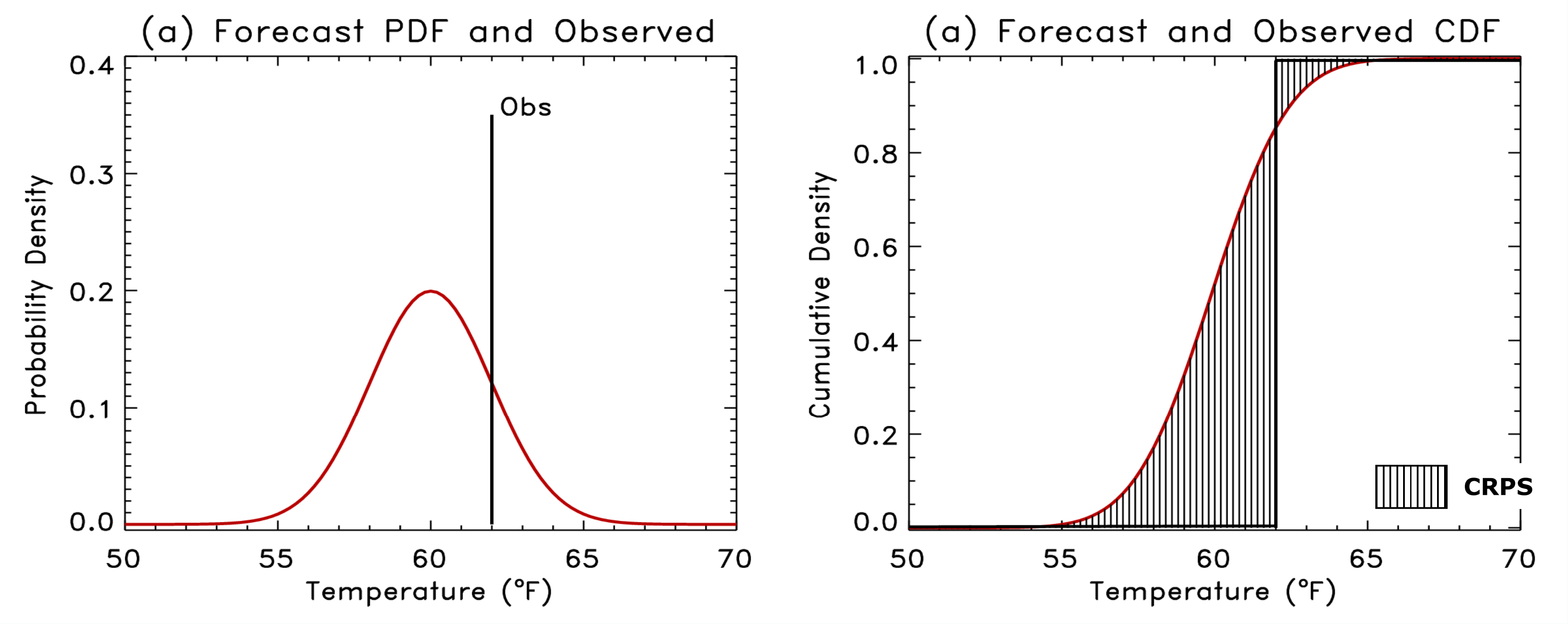
RPS=0.05
not very sharp,
slightly biased

RPS=0.08
accurate,
but not sharp

R. Hagedorn, 2007

Continuous Rank Probability Score

$$CRPS(P, x_a) = \int_{-\infty}^{\infty} \left[P(x) - P_a(x) \right]^2 dx$$

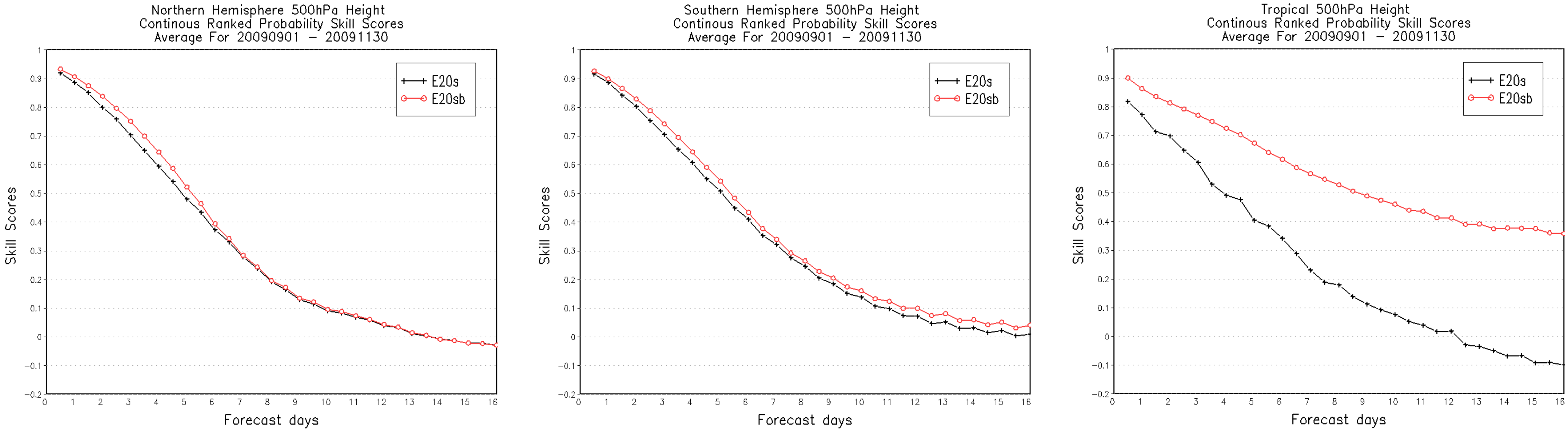


- Area difference between CDF observation and CDF forecast
- Defaults to MAE for deterministic forecast
- Flexible, can accommodate uncertain observations

Continuous Rank Probability Skill Score

$$RPSS = \frac{\overline{RPS} - \overline{RPS}_{reference}}{0 - \overline{RPS}_{reference}} = 1 - \frac{\overline{RPS}}{\overline{RPS}_{reference}}$$

Example:
500hPa CRPS of operational GFS (2009; black) and new implementation (red)



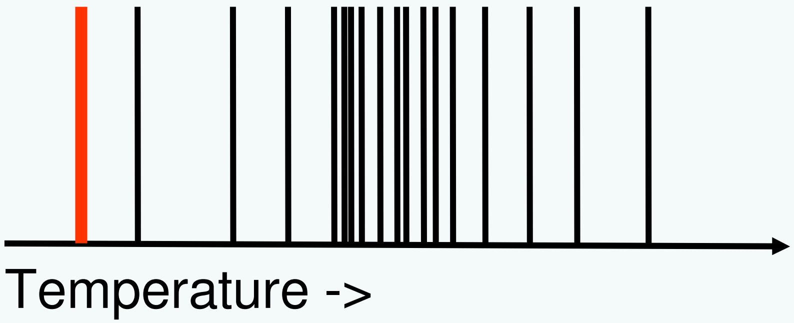
Courtesy of Y. Zhu (EMC/NCEP)

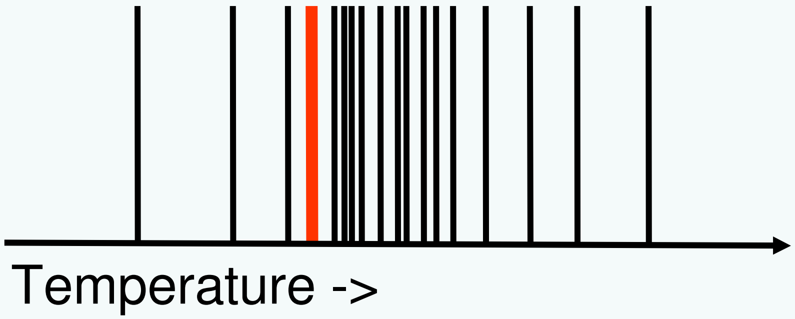
Rank Histogram

- Do the observations statistically belong to the distributions of the forecast ensembles? (consistent degree of ensemble dispersion)
- Diagnose the average spread of an ensemble compared to observations
- Computation: Identify rank of the observation compared to ranked ensemble forecasts
- Assumption: observation equally likely to occur in each of n+1 bins.

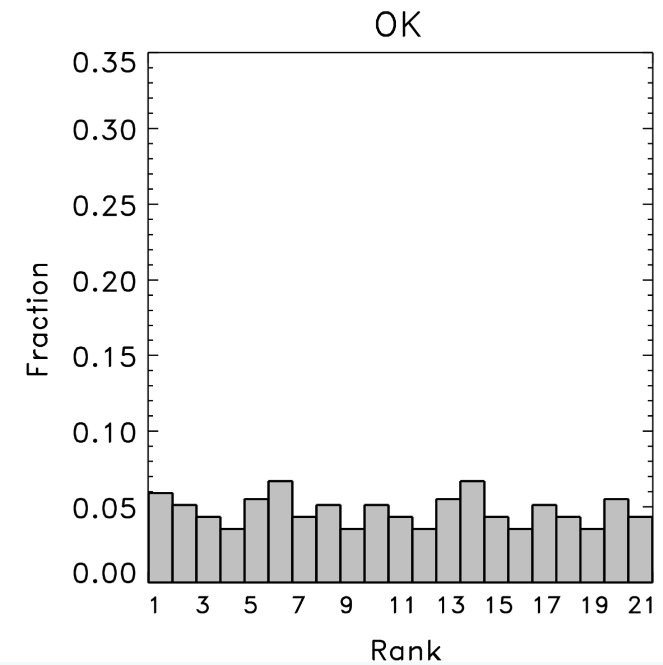
Rank Histogram (Talagrand Diagram)

- Rank Histograms asses whether the ensemble spread is consistent with the assumption that the observations are statistically just another member of the forecast distribution
  - Check whether observations are equally distributed among predicted ensemble
  - Sort ensemble members in increasing order and determine where the observation lies with respect to the ensemble members

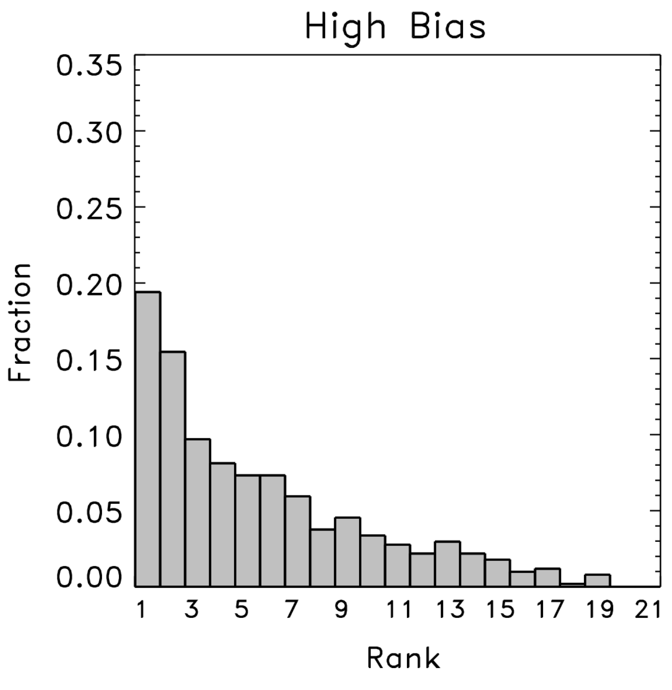

Rank 1 case


Rank 4 case

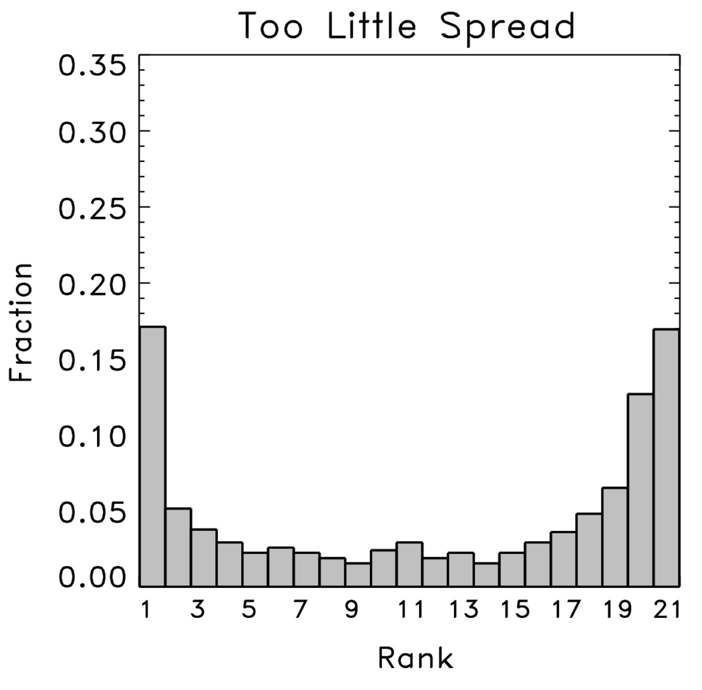
Rank Histograms



OBS is indistinguishable from any other ensemble member



OBS is too often below the ensemble members (biased forecast)



OBS is too often outside the ensemble spread

A uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable (see also: T. Hamill, 2001, MWR)

Comments on Rank Histograms

- Not a real verification measure
- Quantification of departure from flatness

$$RMSD = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(s_k - \frac{M}{N+1}\right)^2}$$

where RMSD is the root-mean-square difference from flatness, expressed as number of cases, *M* is the total sample size on which the rank histogram is computed, *N* is the number of ensemble members, and *sk* is the number of occurrences in the *k*th interval of the histogram.

Definition of a proper score

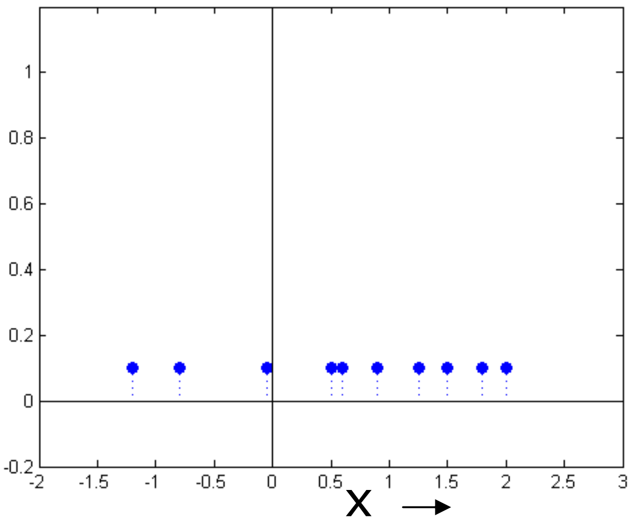
- “Consistency” with your true belief is one of the characteristics of a good forecast
- Some scoring rules encourage forecasters to be inconsistent, e.g. some scores give better results when a forecast closer to climatology is issued rather than the actual forecast (e.g. reliability)
- Scoring rule is *strictly proper* when the best scores are obtained if and only if the forecasts correspond with the forecaster’s judgement (true belief)
- Examples of proper scores are the Brier Score or RPS

Estimating Probabilities from ensembles

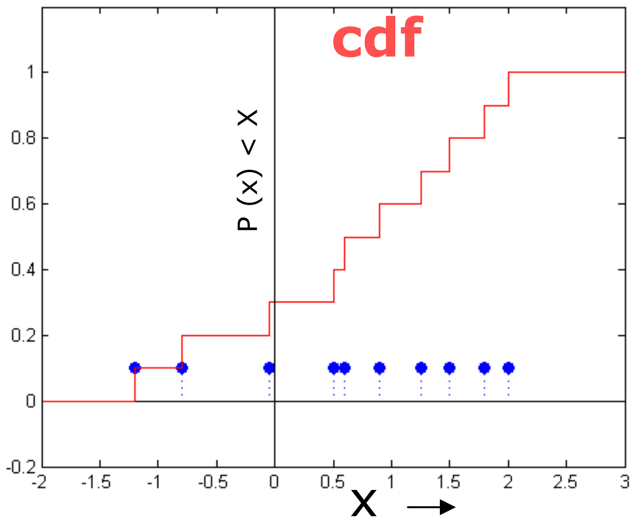
- PDFs and CDFs estimates are interpretations of the Ensemble Prediction System outputs
- Approaches
  - Discrete
  - Histograms
  - Continuous (parametric and nonparametric)

Ensemble Distribution

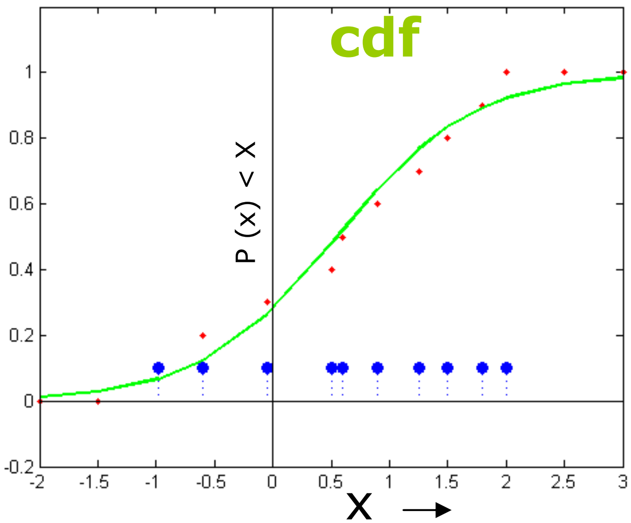
10 Members (1/10 likelihood)



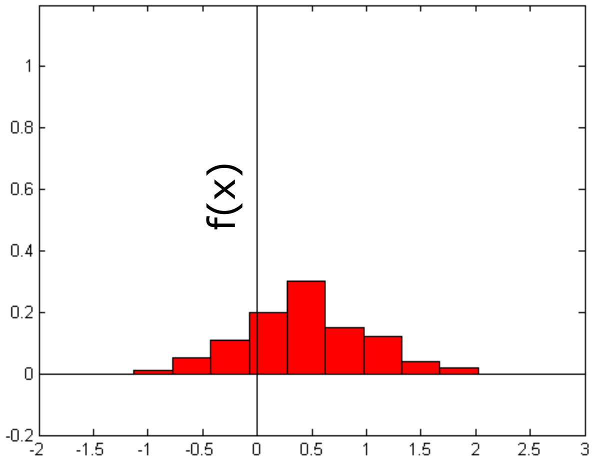
Discrete



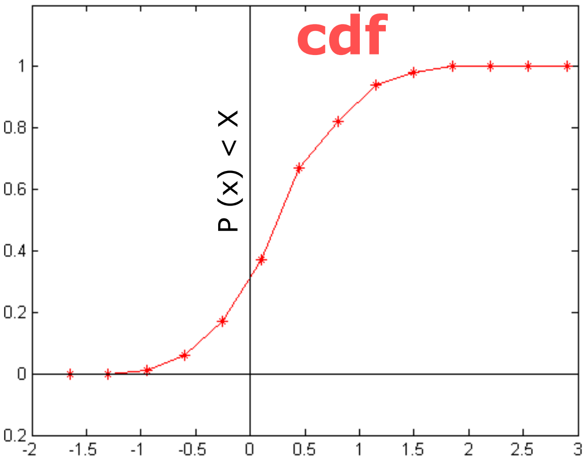
Continuous (Logistic fit)



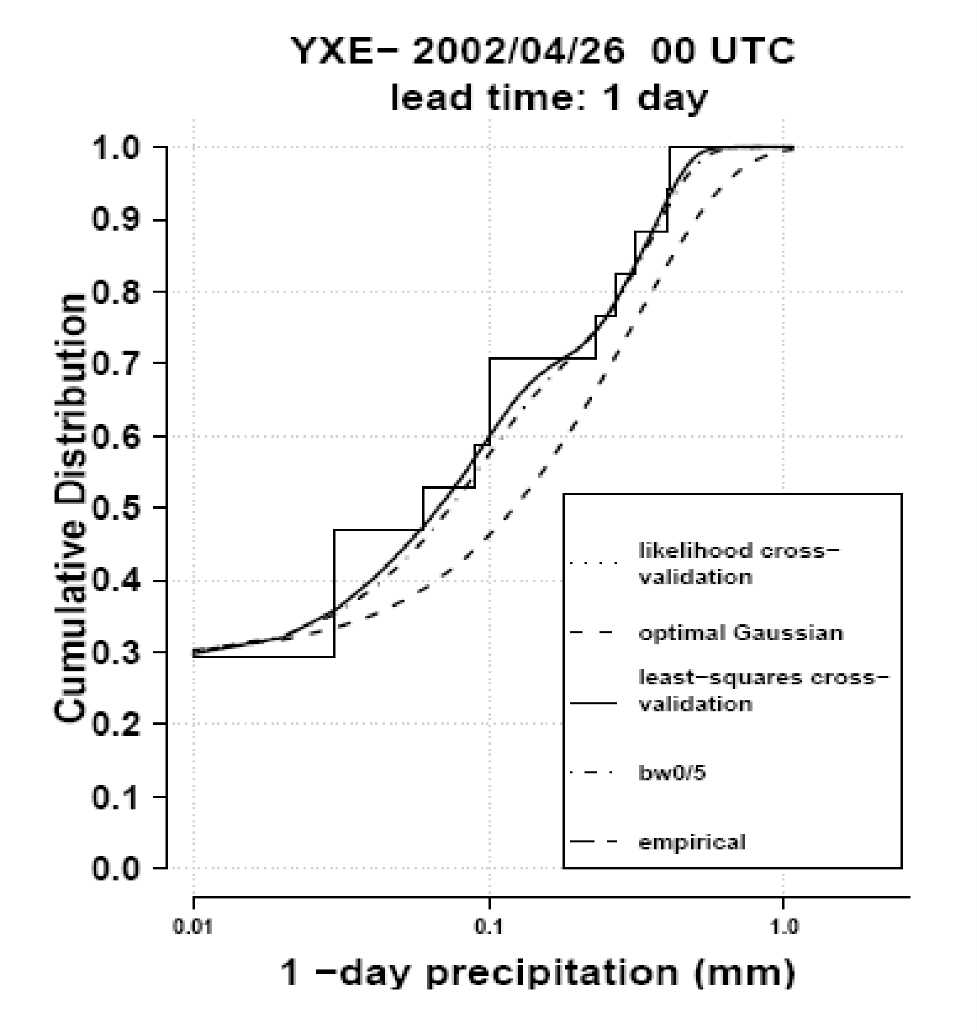
Histogram



Non-parametric



Example of discrete and fitted cdf



From L. Wilson (EC)

Continuous Density fitting uses

- For extreme event prediction, to estimate centile thresholds.
- Assists with the ROC computation
- Simple to compare ensembles with different numbers of members
- Several approaches to estimate “density” from observational data (e.g., Silverman, 1986)
- One of those is the Kernel approach. Advantages: Non-parametric; amount of smoothing determined by the bandwidth; Gaussian kernels fine for unbounded variables; Gamma kernels for precipitation

Definition of Kernel

- Kernel is a weighting function satisfying the following two requirements:

1) $\int_{-\infty}^{\infty} K(u)du = 1;$ $\longrightarrow$ To ensure estimation results in a PDF

2) $K(-u) = K(u); \quad \forall u$ $\longrightarrow$ To ensures that the average of the corresponding distribution is equal to that of the sample used

- Examples:

Uniform

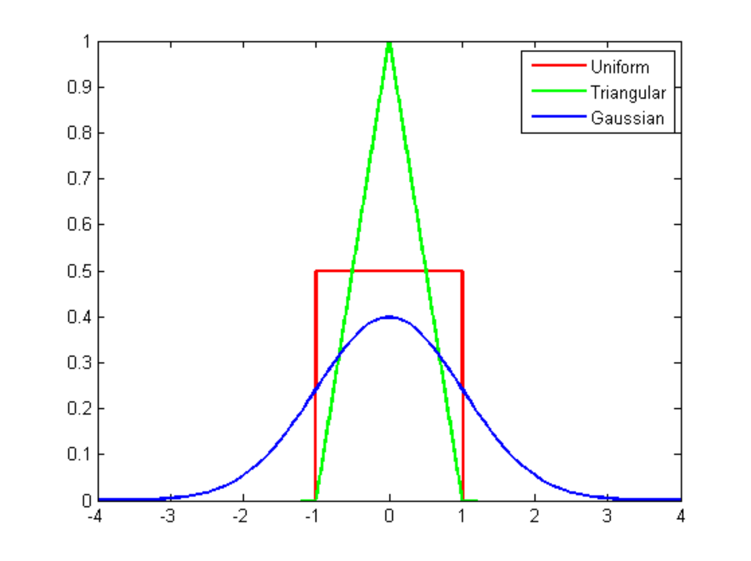
$$K(u) = \frac{1}{2}\mathbf{1}_{\{|u|\le 1\}}$$

Triangular

$$K(u) = (1 - |u|)\mathbf{1}_{\{|u|\le 1\}}$$

Gaussian

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$



- Very often, the kernel is taken to be a Gaussian function with mean zero and variance 1. In this case, the density is controlled by one smoothing parameter h (bandwidth)

$$u = \frac{x - x_i}{h}$$

where x_i is an independent sample of a random variable f; thus the density is given as:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Calibration

- Forecasts of Ensemble Prediction System are subject to forecast bias and dispersion errors
- Calibration aims at removing such known forecast deficiencies, i.e. to make statistical properties of the raw EPS forecasts similar to those from observations
- Calibration is based on the behaviour of past EPS forecast distributions, therefore needs a record of historical prediction-observation pairs
- Calibration is particularly successful at station locations with long historical data records
- A growing number of calibration methods exist and are becoming necessary to process multi-model ensembles

Calibration methods for EPS’s

- Systematic error correction
- Multiple implementation of deterministic MOS
- Ensemble dressing
- Bayesian model averaging
- Non-homogenous Gaussian regression
- Logistic regression
- Analog method

Examples of systematic errors

$$d_{SE}(f_t,o_t)=\bar{d}(f_t)-\bar{d}(o_t)$$

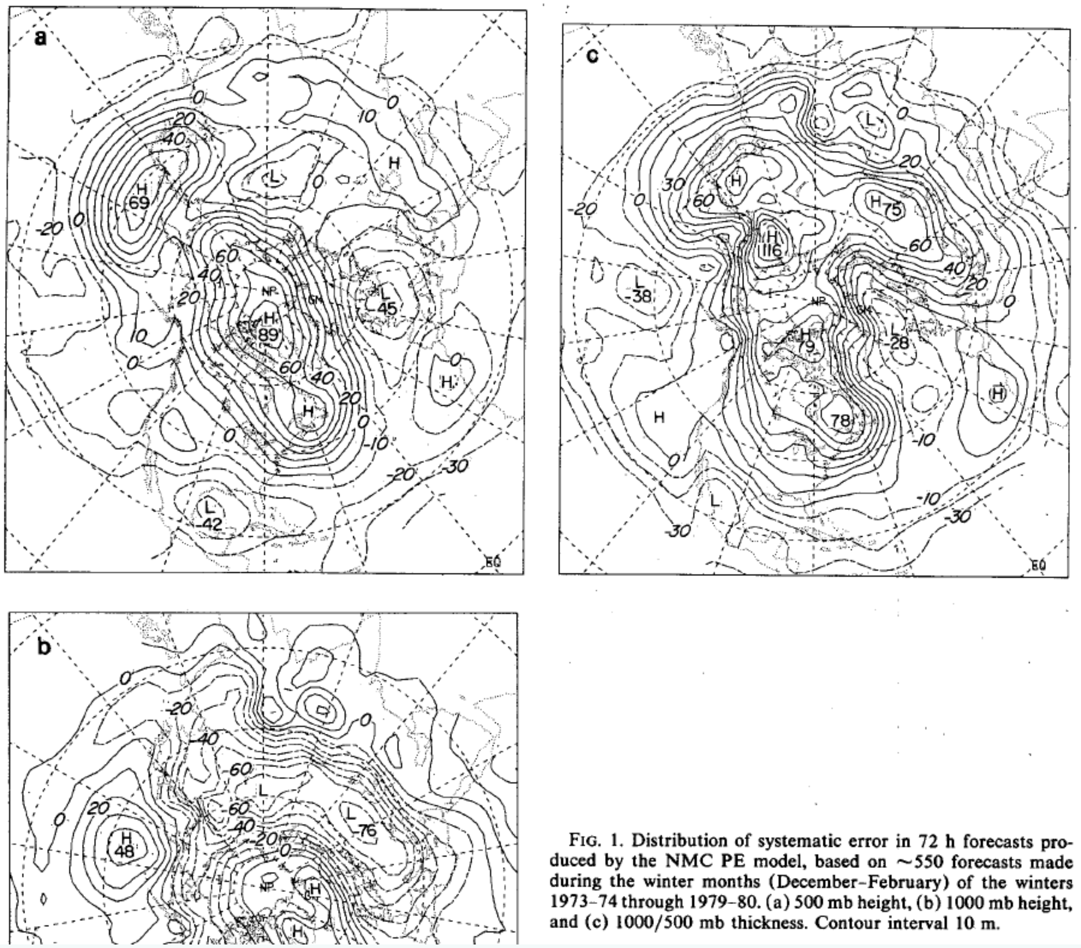
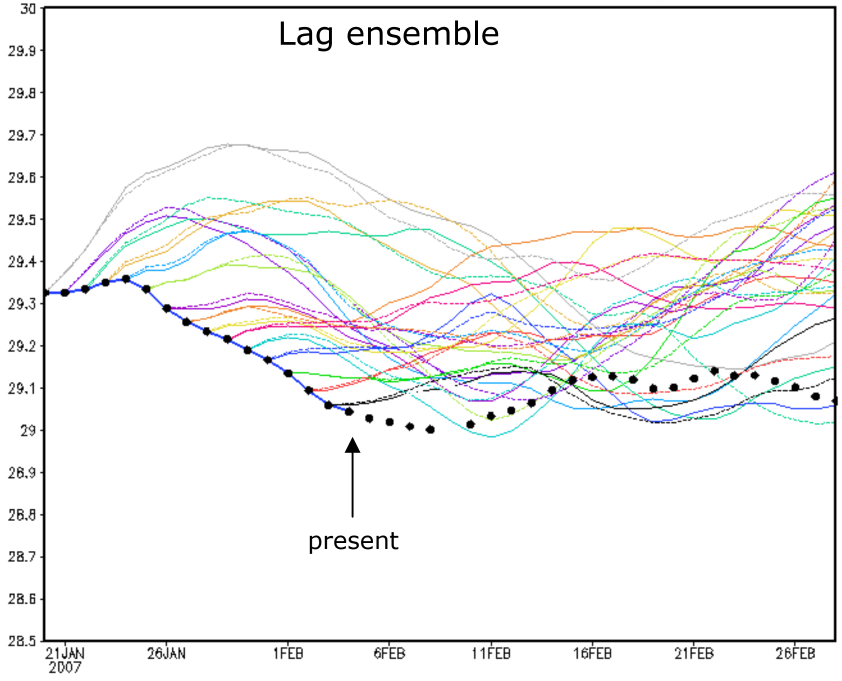


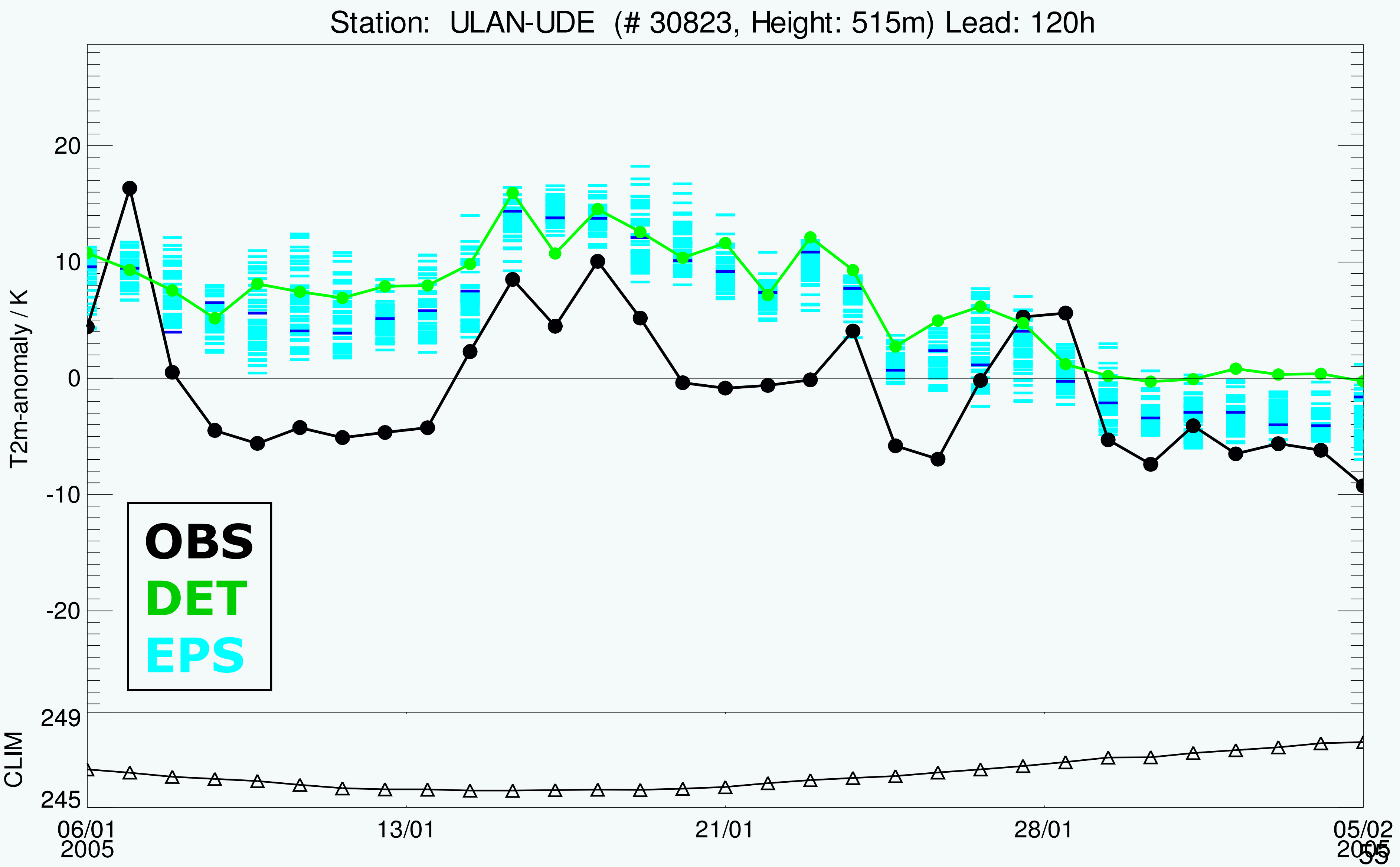
FIG. 1. Distribution of systematic error in 72 h forecasts produced by the NMC PE model, based on ~550 forecasts made during the winter months (December–February) of the winters 1973–74 through 1979–80. (a) 500 mb height, (b) 1000 mb height, and (c) 1000/500 mb thickness. Contour interval 10 m.

Nino3.4 Forecasts and analysis



Note “Warm” tendency of the model

Example of Bias



R. Hagedorn, 2007

Bias correction

- As a simple first order calibration a bias correction can be applied using:

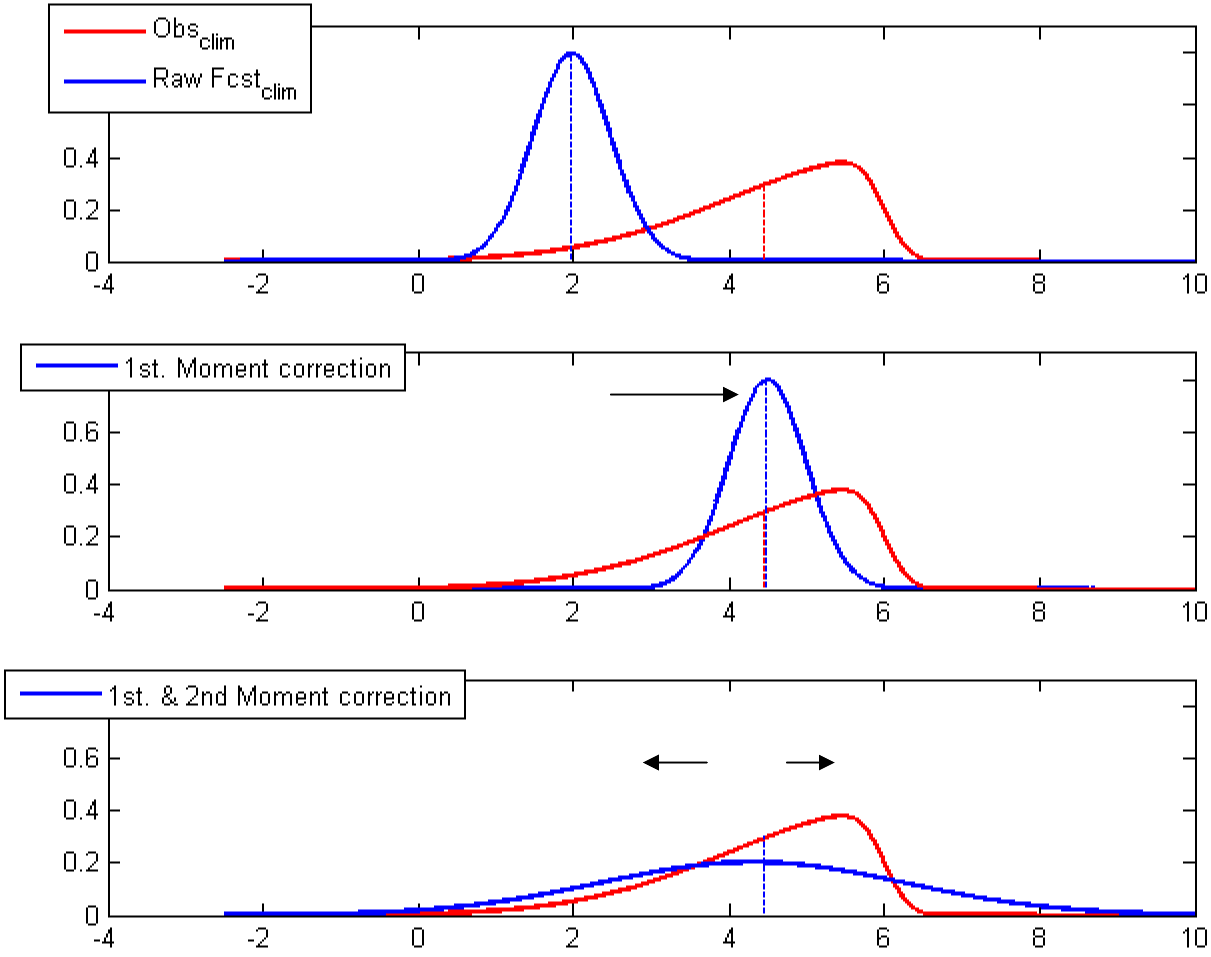
$$c = \frac{1}{N}\sum_{i=1}^{N}\overline{e}_i - \frac{1}{N}\sum_{i=1}^{N}o_i$$

with: $\overline{e}_i$ = ensemble mean of the ith forecast
o_i = value of ith observation
N = number of observation-forecast pairs

- This correction factor is applied to each ensemble member, i.e. spread is not affected

- Particularly useful/successful at locations with features not resolved by model and causing significant bias

Illustration of first and second moment correction



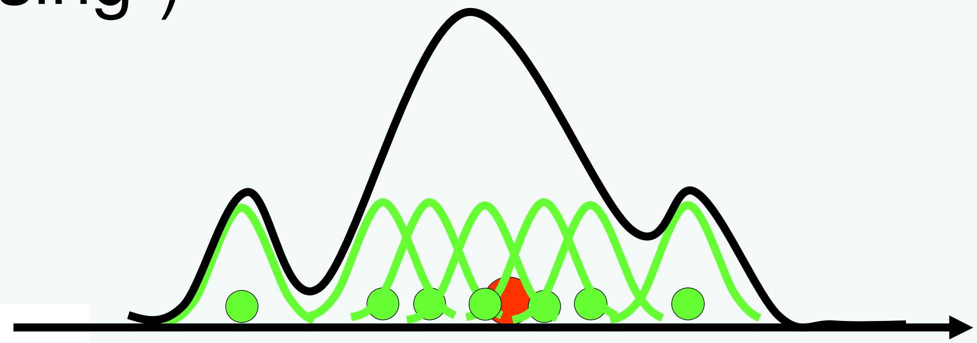
Multiple implementation of det. MOS

- A possible approach for calibrating ensemble predictions is to simply correct each individual ensemble member according to its deterministic model output statistic (MOS)
- **BUT**: this approach is conceptually inappropriate since for longer lead-times the MOS tends to correct towards climatology
  - all ensemble members tend towards climatology with longer lead-times
  - decreased spread with longer lead-times
  - in contradiction to increasing uncertainty with increasing lead-times

(Further reading on this problem: Vannitsem (2009), QJRMS)

Ensemble dressing

- Define a probability distribution around each ensemble member ("dressing")



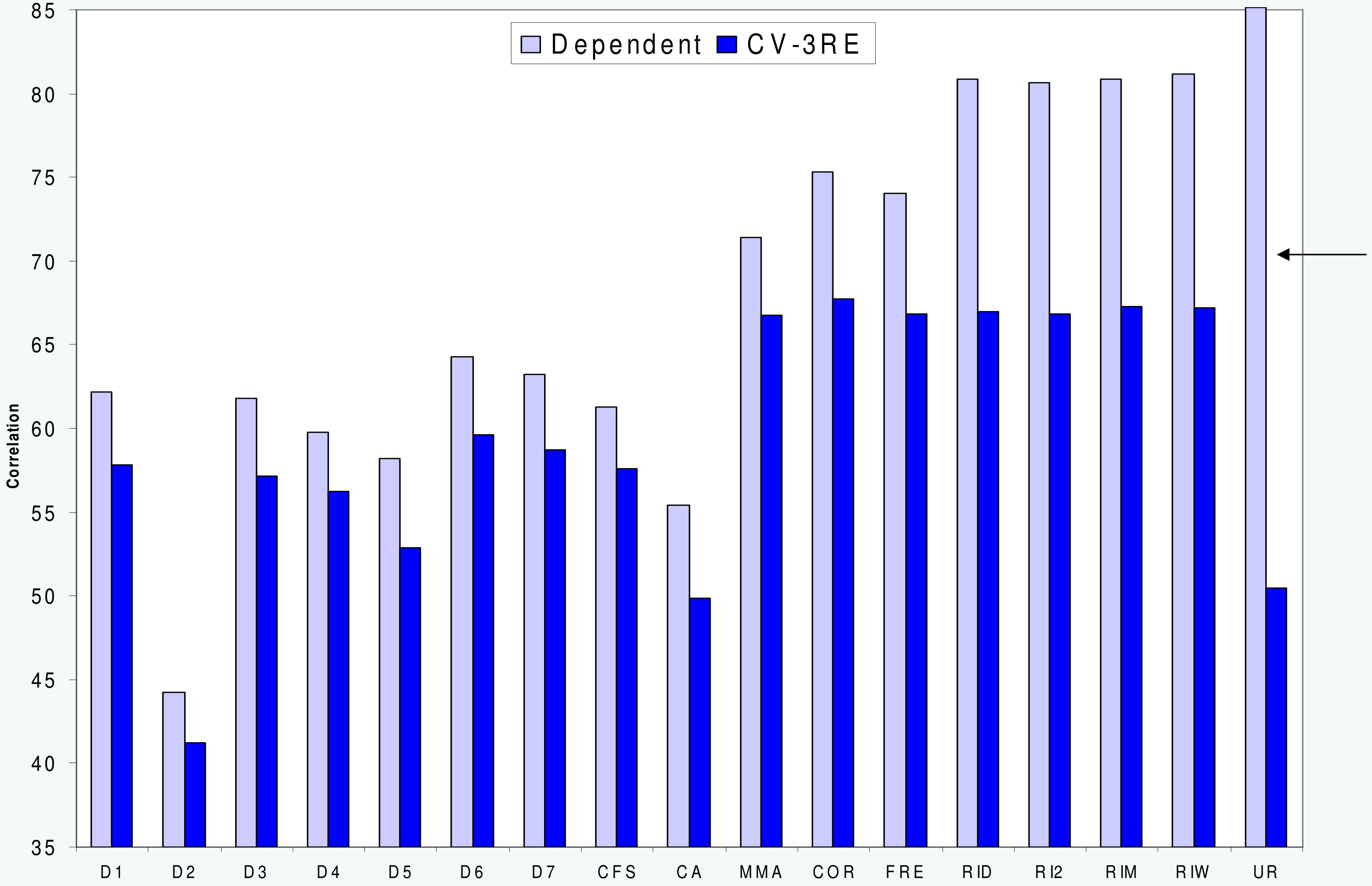
- A number of methods exist to find appropriate dressing kernel ("best-member" dressing, "error" dressing, "second moment constraint" dressing, etc.)
- Average the resulting n_{ens} distributions to obtain final pdf

R. Hagedorn, 2007

Training datasets

- All calibration methods need a training dataset (hindcast), containing a large number historical pairs of forecast-observation fields
  - A long training dataset is preferred to more accurately determine systematic errors and to include past extreme events
  - Common approaches: a) generate a sufficiently long hindcast and freeze the model; b) compute a systematic error but the model is not frozen.
- For research applications often only one dataset is used to develop and test the calibration method. In this case it is crucial to carry out cross-validation to prevent “artificial” skill.

Effect of cross-validation on multi-model combination methods



Artificial skill

References

Atger, F., 1999: The skill of Ensemble Prediction Systems. Mon. Wea. Rev. 127, 1941-1953.

Hamill, T., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon. Wea. Rev., 129, 550-560.

Silverman, B.W., 1986: Density Estimation for Statistical and Data Analysis, Chapman and Hall Ltd. 175

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2002: Probability and ensemble forecasts. In: Environmental Forecast Verification: A practitioner's guide in atmospheric science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, pp.137-164.

Vannitsem, S., 2009: A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, vol. 135, issue 644, pp. 1801-1815.

Wilks, D., 1995: Statistical Methods in the Atmospheric Sciences. Academic Press, 464pp.

Internet sites with more information:

http://wwwt.emc.ncep.noaa.gov/gmb/ens/index.html

http://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts

http://www.ecmwf.int/newsevents/training/meteorological_presentations/MET_PR.html