April 16th, 2014

Ensemble Forecasting and their Verification

Malaquías Peña

Environmental Modeling Center, NCEP/NOAA

Material comprises Sects. 6.6, 7.4 and 7.7 in Wilks (2nd Edition). Additional material and notes from Zoltan Toth (NOAA) and Yuejian Zhu (NOAA), Renate Hagedorn (ECMWF), and Laurence Wilson (EC)

Outline

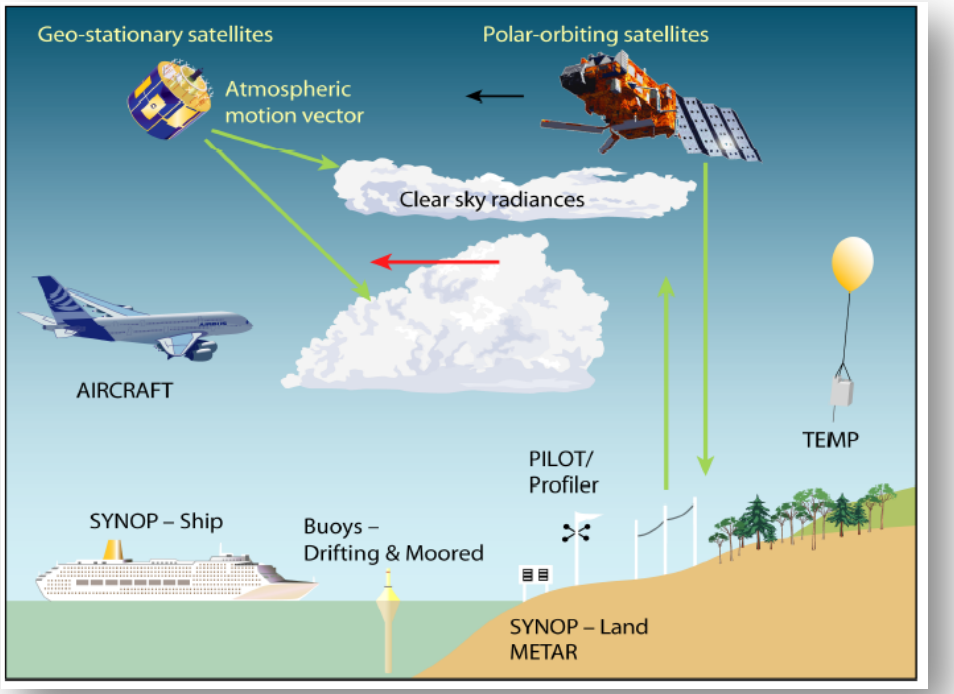
1. Key concepts
   - Uncertainty in NWP systems
   - Ensemble Forecasting
2. Probabilistic forecast verification
   - Attributes of forecast quality
   - Performance metrics
3. Post-processing ensembles

Uncertainties in NWP

Initial state of the atmosphere

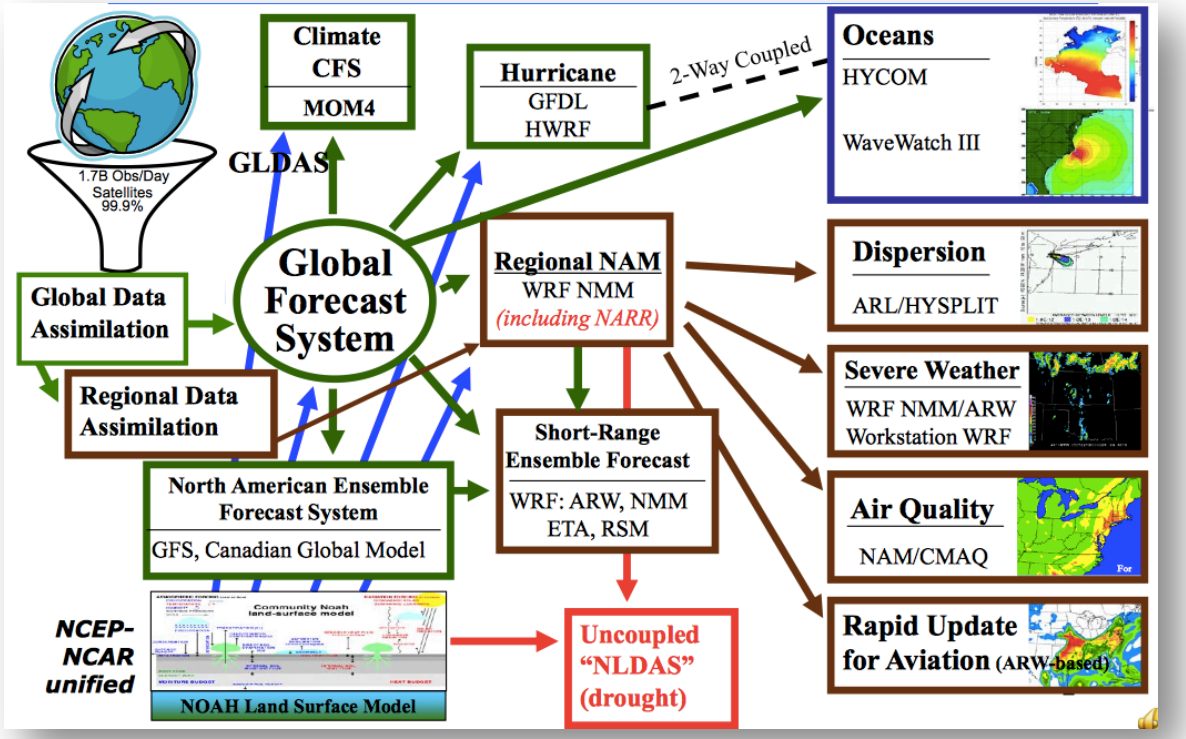
Errors in the observations

- Precision errors
- Bias in frequency of measurements,
- **Representativeness** errors,
- Reporting errors,
- Random errors
- Conversion errors, etc.



Uncertainties in NWP

Initial state of the atmosphere



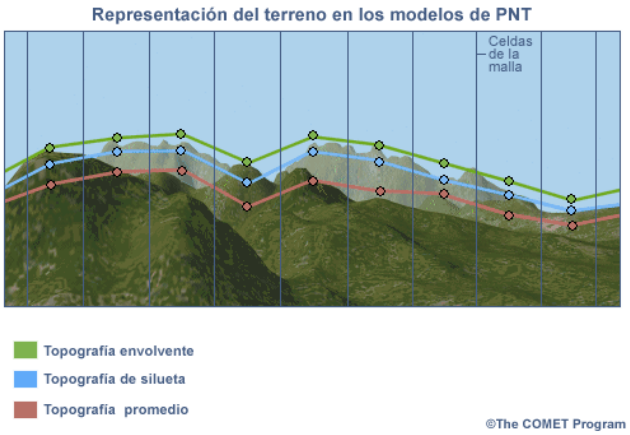
Bias in the first guess

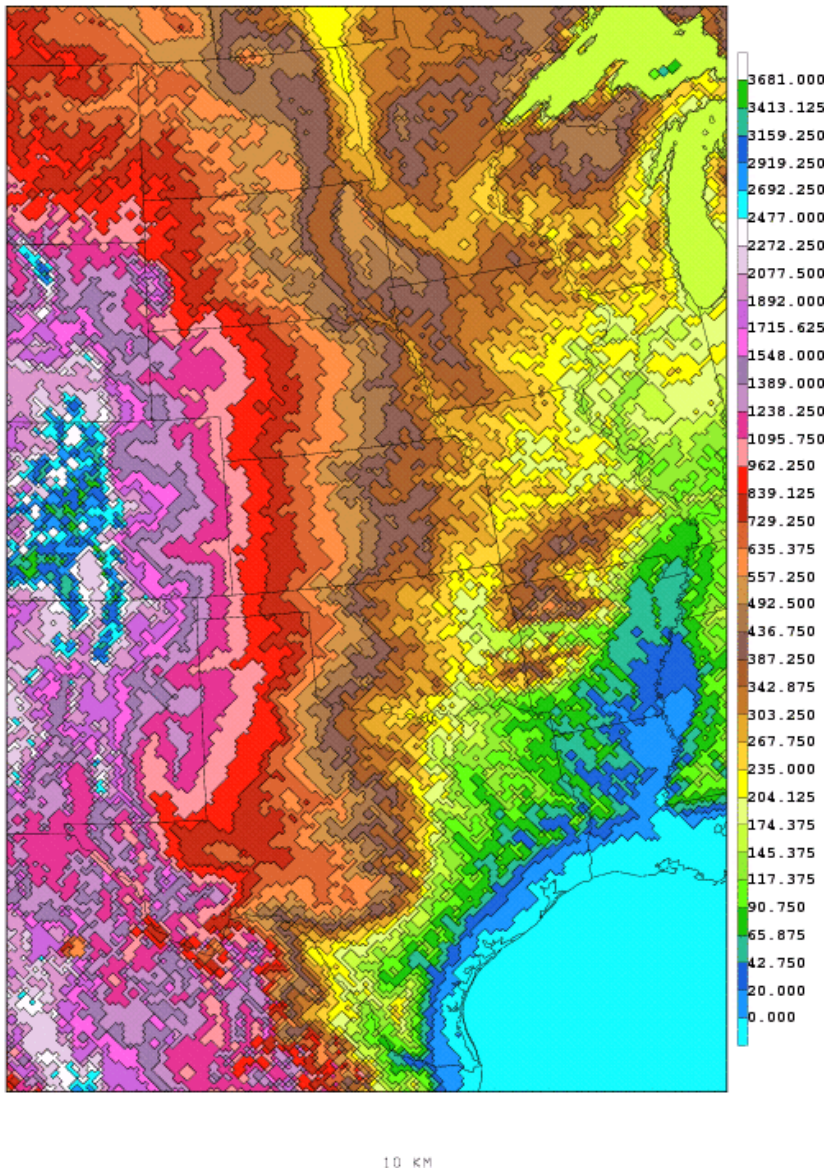
- Deficient representation of relevant physical processes
- Truncation errors
- Limited approach for data QC: incorrect good vs bad data discrimination
- Computer program bugs!

Uncertainties in NWP

Representing the physical processes in the atmosphere

Insufficient spatial resolution, truncation errors in the dynamical equations, limitations in parameterization, etc.



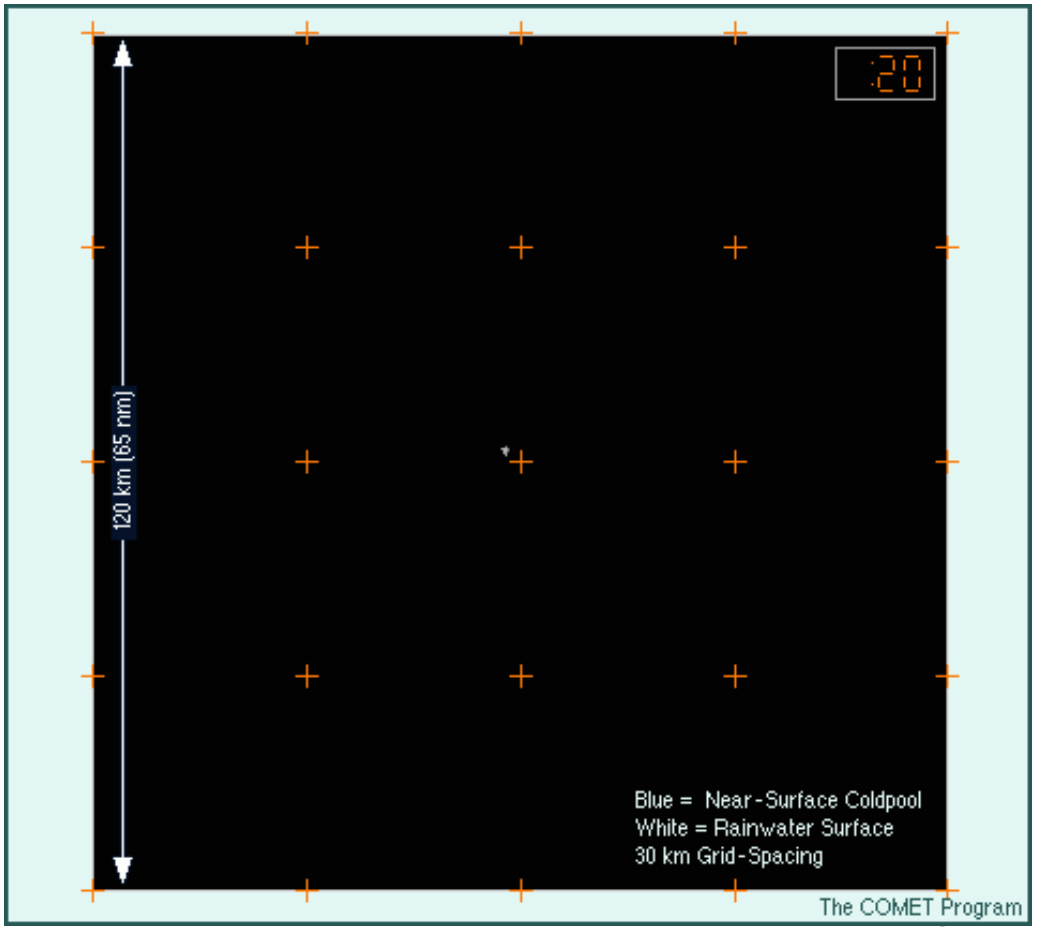


Uncertainties in NWP

Model errors: Uncertainty describing the evolution of the atmosphere

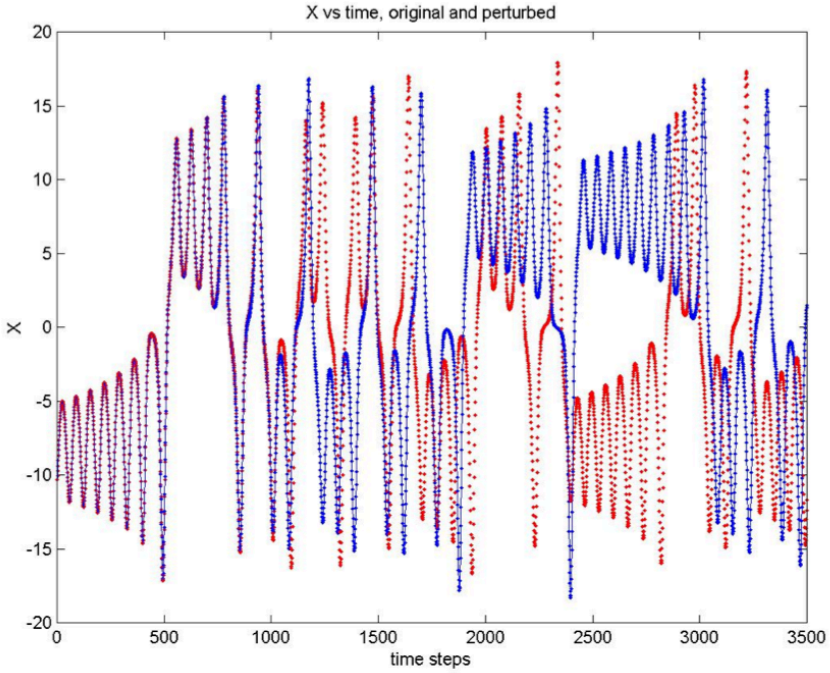
Sources: ad hoc parameterization, average errors, coding errors!, bias in frequency of initialization, etc.

Any process occurring between grid points will go unnoticed by the model



Uncertainties in NWP

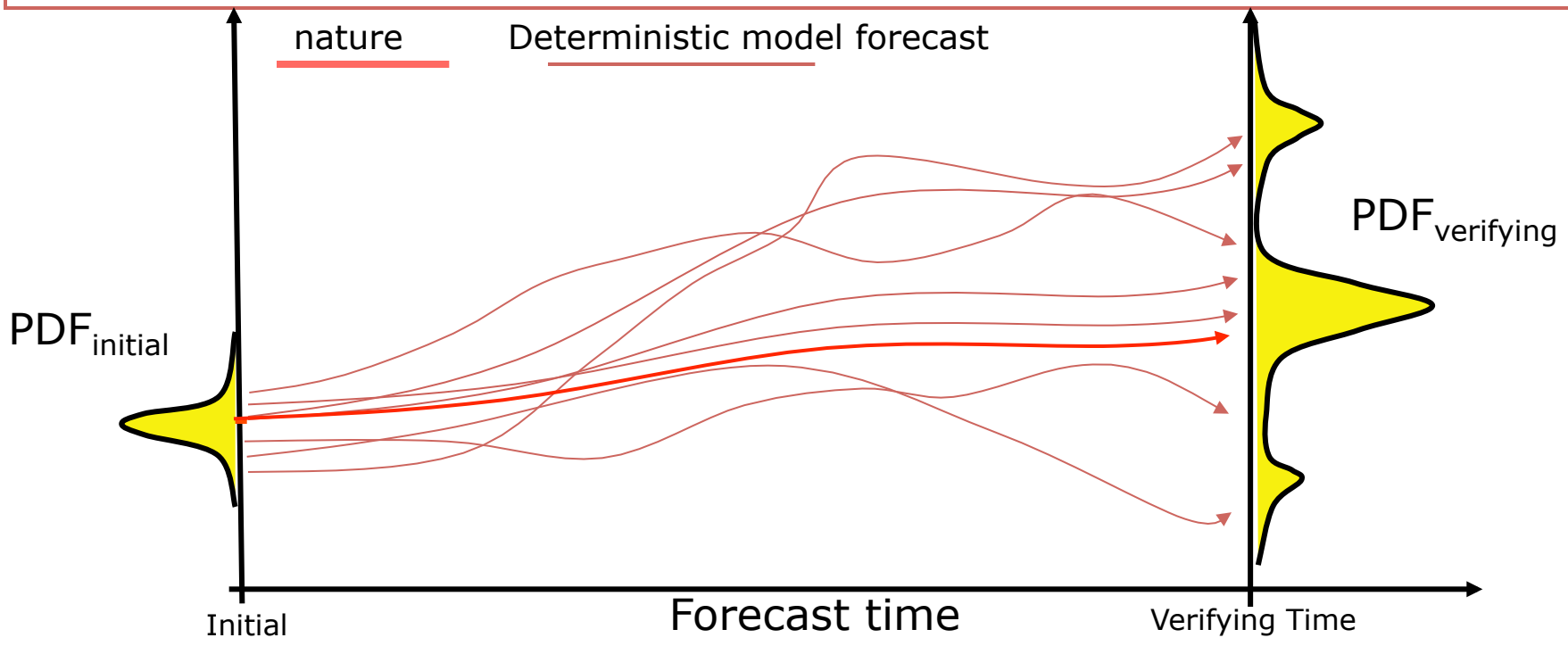
Butterfly effect: The solution of the equations in the NWP systems are sensitive to initial conditions



Even if initial errors were largely reduced, any small deviation will render completely different solutions after several integration time steps.

Ensemble forecast

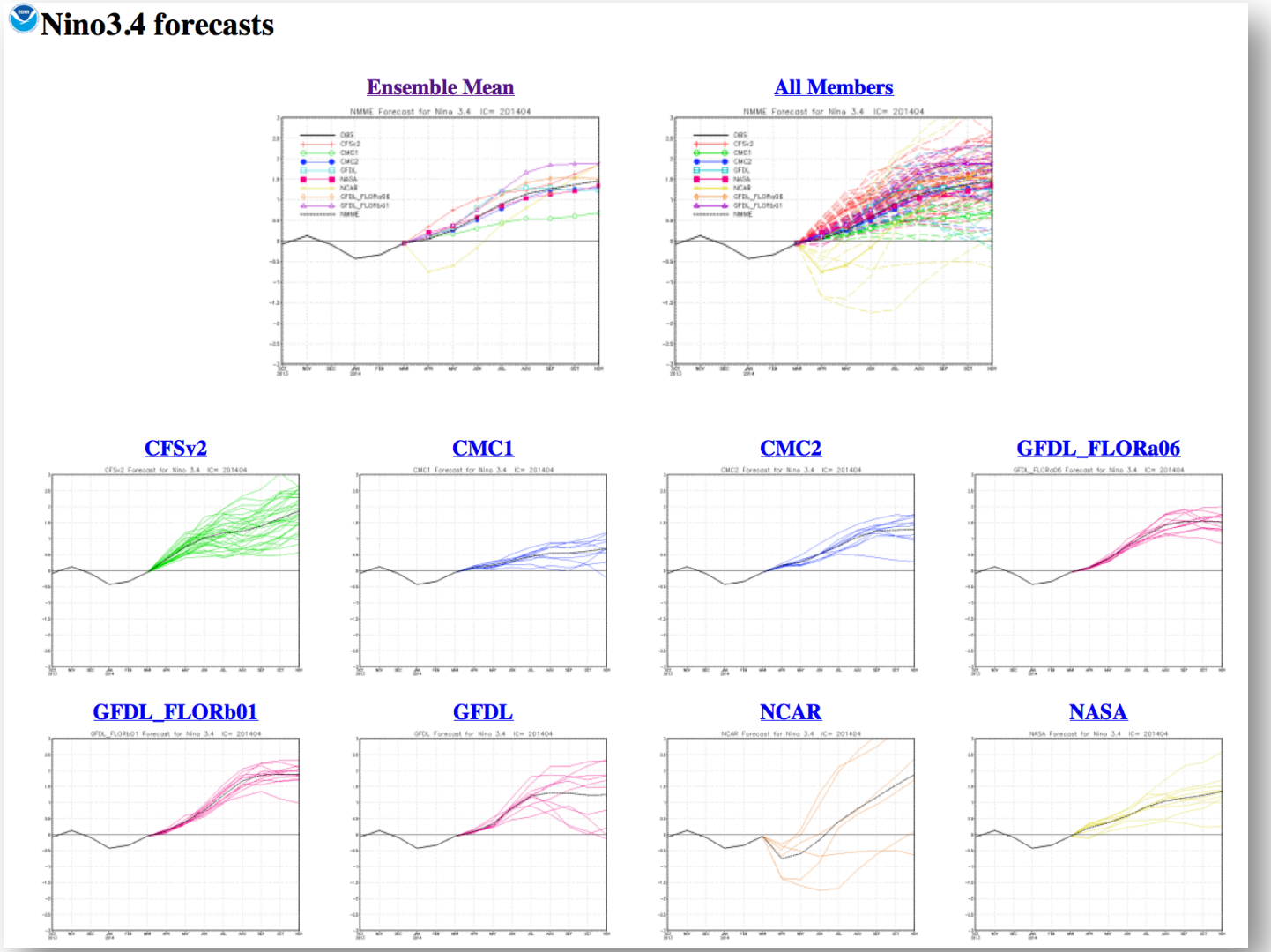
- A collection of two or more forecasts verifying at the same time
- Ensemble forecasting: propagate into de future the probability distribution function reflecting the uncertainty associated with initial conditions and model limitations
- Practical approach: running a set of deterministic runs whose initial conditions sample the initial PDF



Visualization of Ensembles

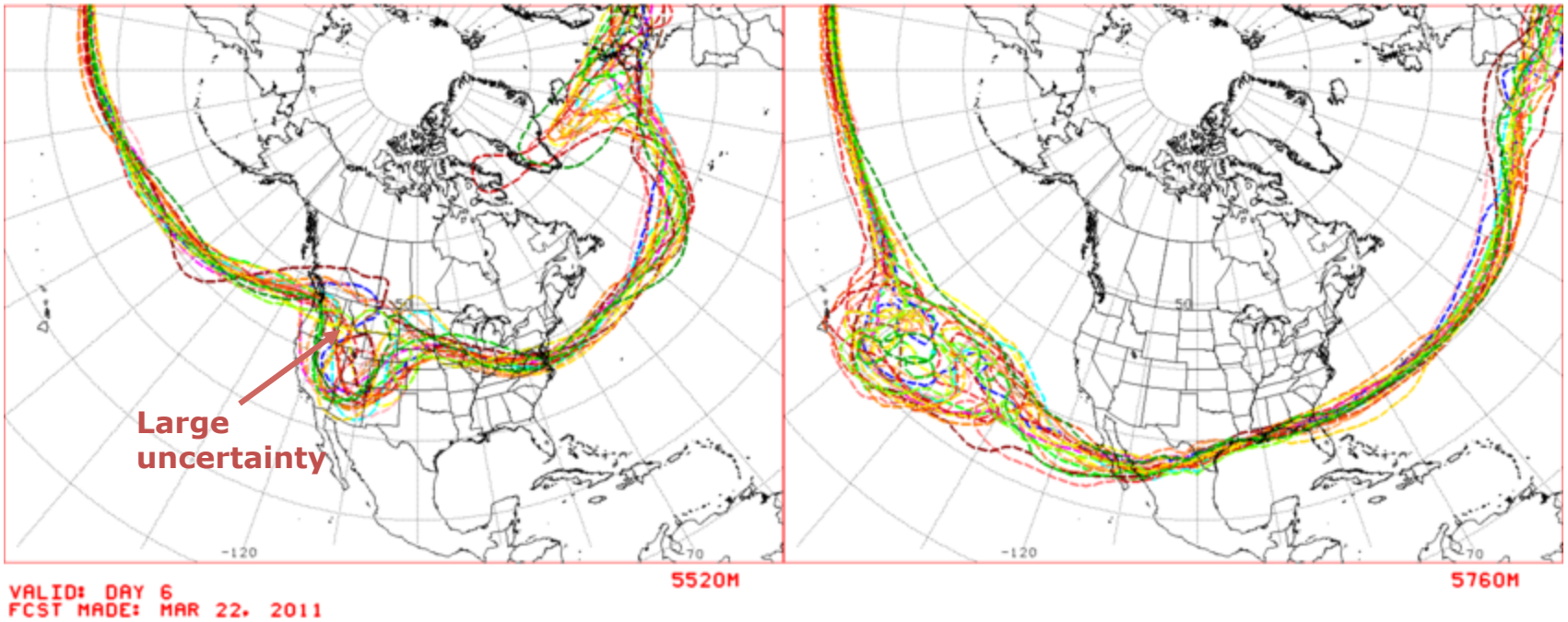
Multi-model Ensemble

NCEP displays Nino 3.4 index forecasts from ensemble seasonal prediction systems.



Visualization of Ensembles

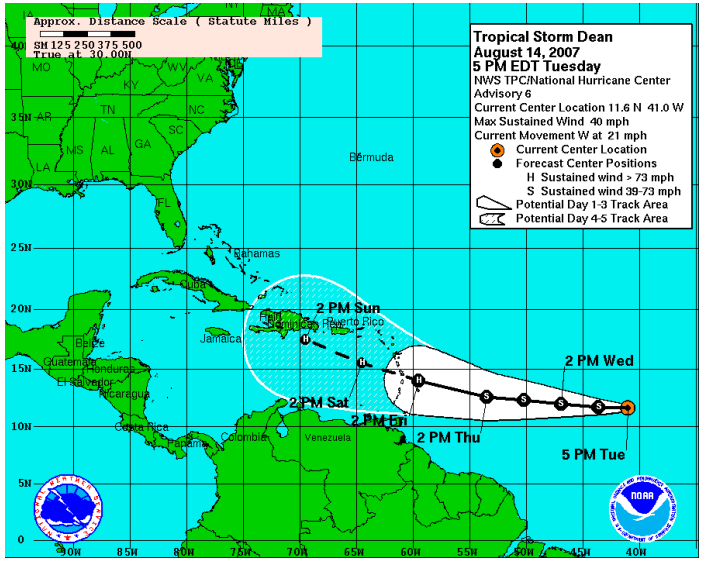
Spaghetti Plots



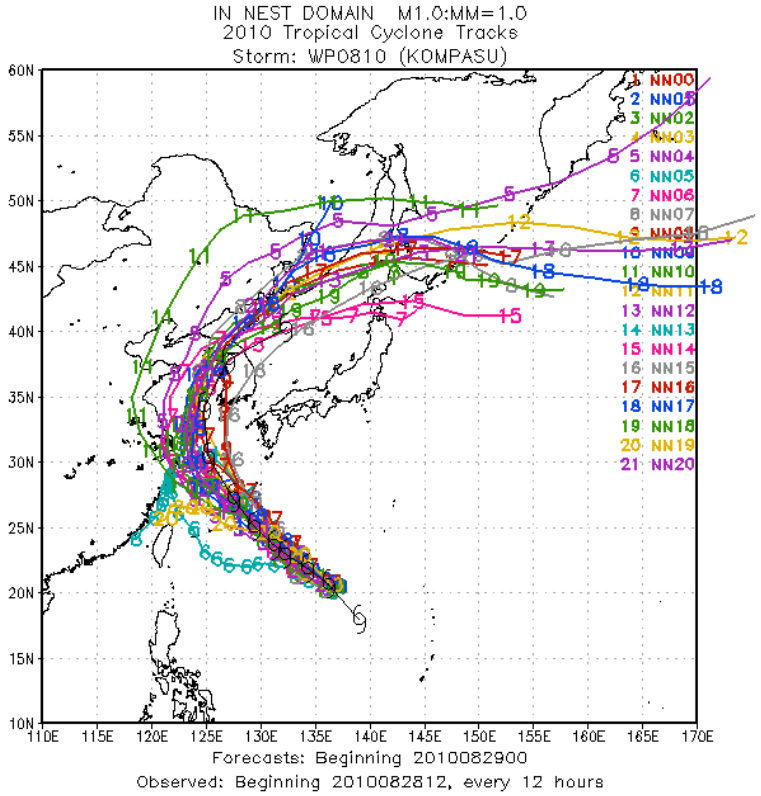
- Contours of ensemble forecasts at specific geopotential heights at 500hPa
- Visualizing the amount of uncertainty among ensemble members
  - High confidence of the forecast in regions where members tend to coincide
- Advantages over mean-spread diagrams: keeps features sharp, allows identifying clustering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

Visualization of Ensembles

Hurricane Tracks



Based on historical official forecast errors over a 5-year sample.

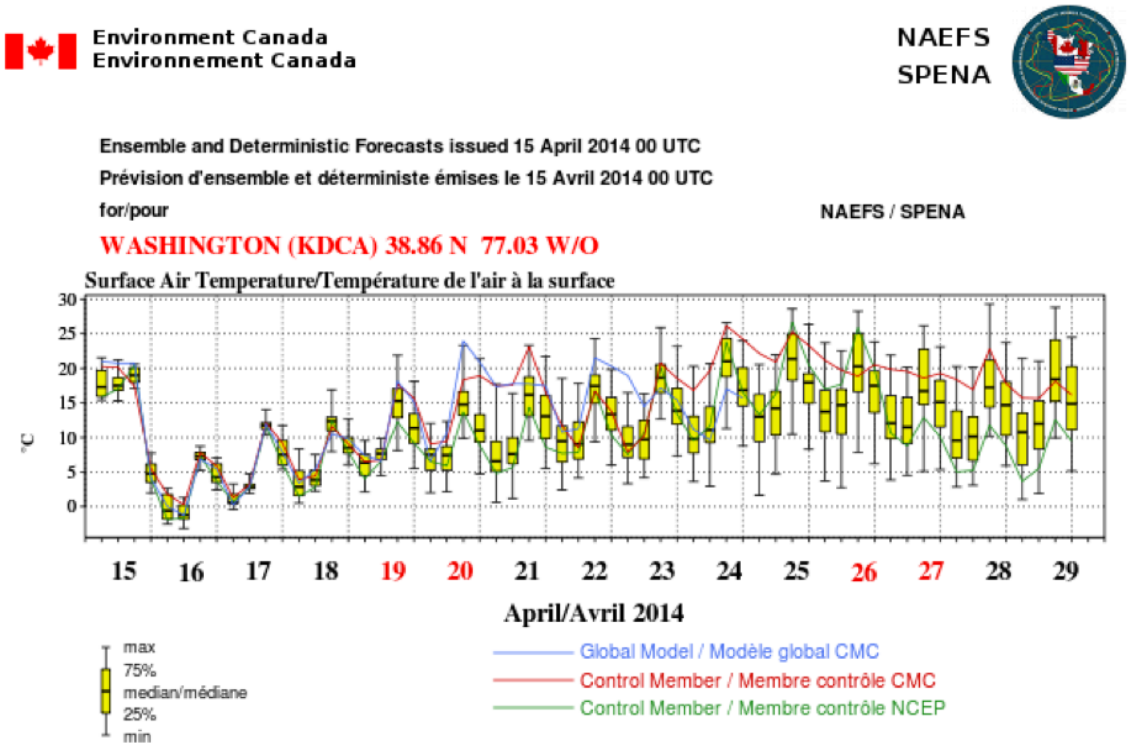


Forecast runs from slightly different initial conditions

Can you tell advantages and disadvantages of each approach?

Visualization of Ensembles

EPS-gram



Quantitative description: The Mean

- Approach to convert from probabilistic to deterministic and simplify the interpretation and the evaluation of ensemble. The PDF described by the ensemble is then reduced into a single number

$$\mu = \sum_{i=1}^{N} x_i p_i$$

Example: Forecast ensemble mean of the geopotential height at a model grid point (λ, θ)

$$\overline{\phi}(\lambda, \theta) = \frac{1}{N} \sum_{i=1}^{N} \phi_i(\lambda, \theta)$$

What assumption was made to replace the mean by an average?

- Average removes short-lived variations retaining slowly-varying patterns
- Median or mode could also work to reduce the full PDF into a single number

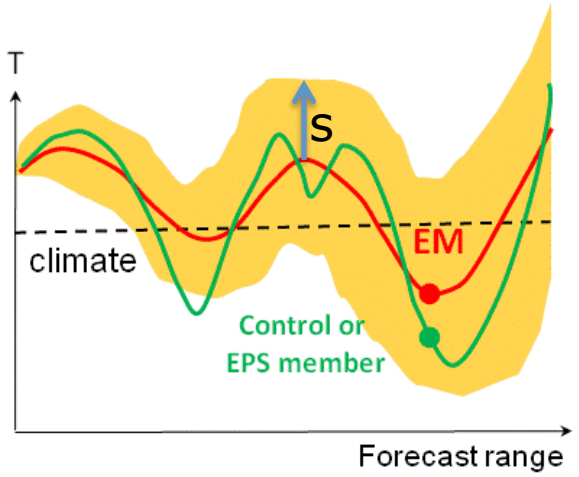
Quantitative description: The Spread

- A measure of the separation among ensemble members. It is the Standard Deviation with respect to the Ensemble Mean
  - In principle, small spread implies less uncertain forecast

$$\sigma = \left(\sum_{i=1}^{N} (x_i - \mu)^2 \, p_i \right)^{1/2}$$

Example: Forecast ensemble spread of the geopotential height at a model grid point (λ, θ)

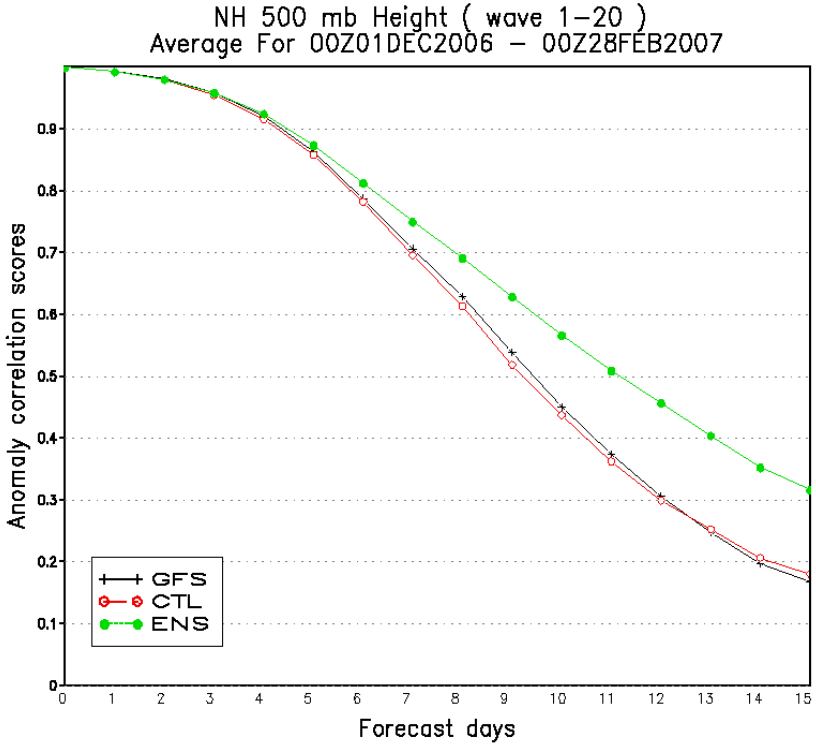
$$S(\lambda,\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} [\phi_i(\lambda,\theta) - \bar{\phi}(\lambda,\theta)]^2 \right)^{1/2}$$



- If the distribution created by 100 members ensemble were normally distributed around the ensemble mean, approximately how many members on average would fall outside the (orange shade in the above schematic) spread line?

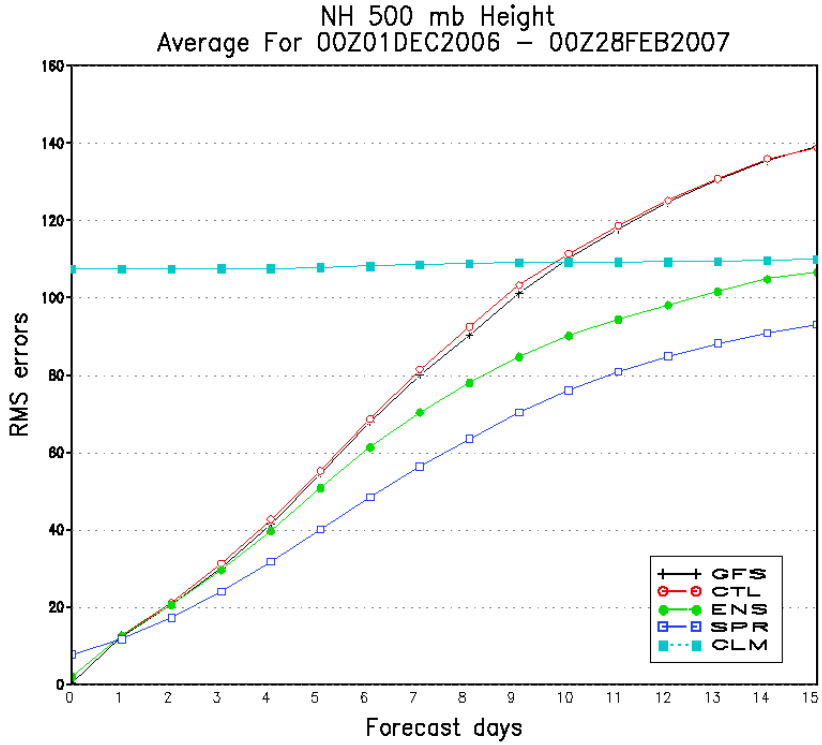
Ensemble Mean Performance

Anomaly Correlation Score



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

Root Mean Square Error



At long leads, the RMSE of the ensemble mean (green) approach climatology (cyan). RMSE of individual forecast members (red and black) grow much faster.

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)

n_t = ensemble size

n_x = number of ensemble members that predict an event x

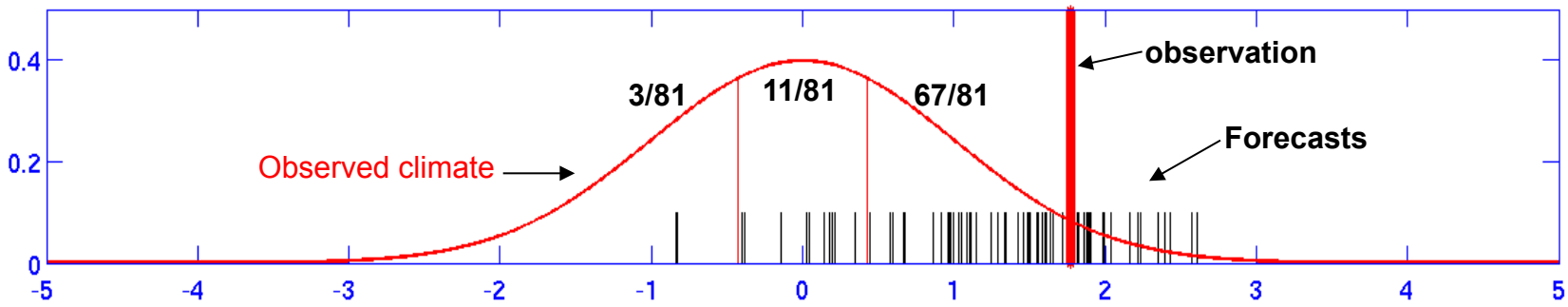
$P(x)$ = probability that event x will happen

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

Quantitative description

Example: An 81-members ensemble forecast is issued to predict the likelihood that a variable will verify in the upper tercile of the historical distribution (variable's climate).

Let's call this event x. Looking at the diagram below we simply count the number of members falling in that tercile.

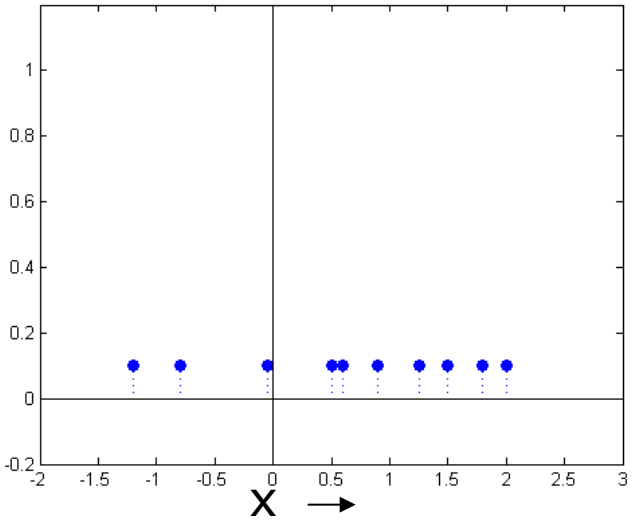


$$P(x) \approx \frac{n_x}{n_t} = \frac{67}{81}$$

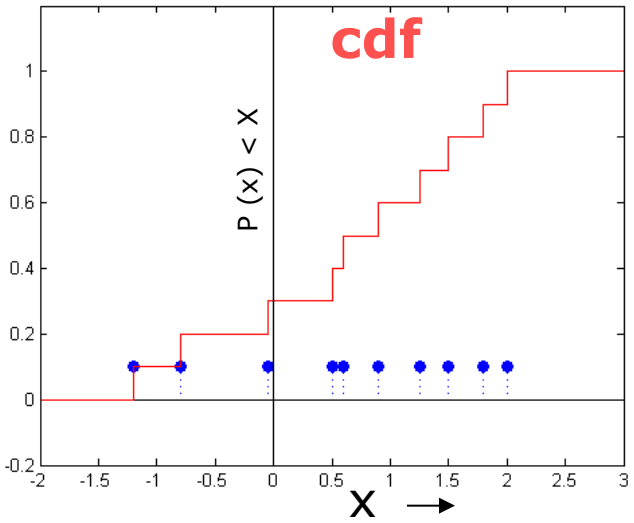
Cumulative Distribution Function

Consider a 10-members ensemble, each with equal (1/10) likelihood of occurrence.

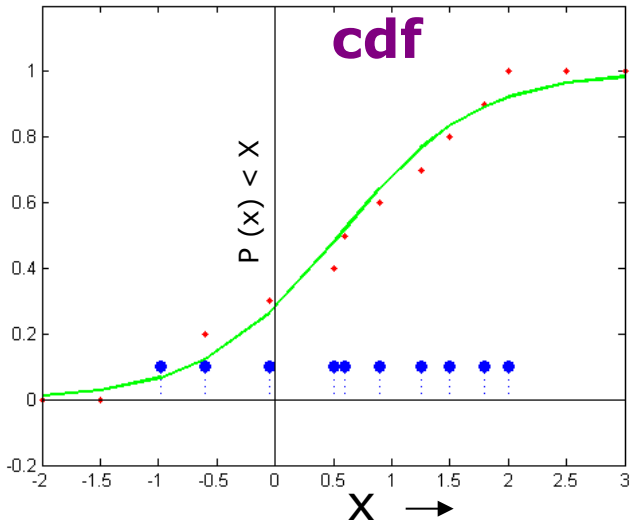
A CDF can be constructed using a discrete (left) or a continuous (right) function



Discrete



Continuous (e.g. Logistic fit)

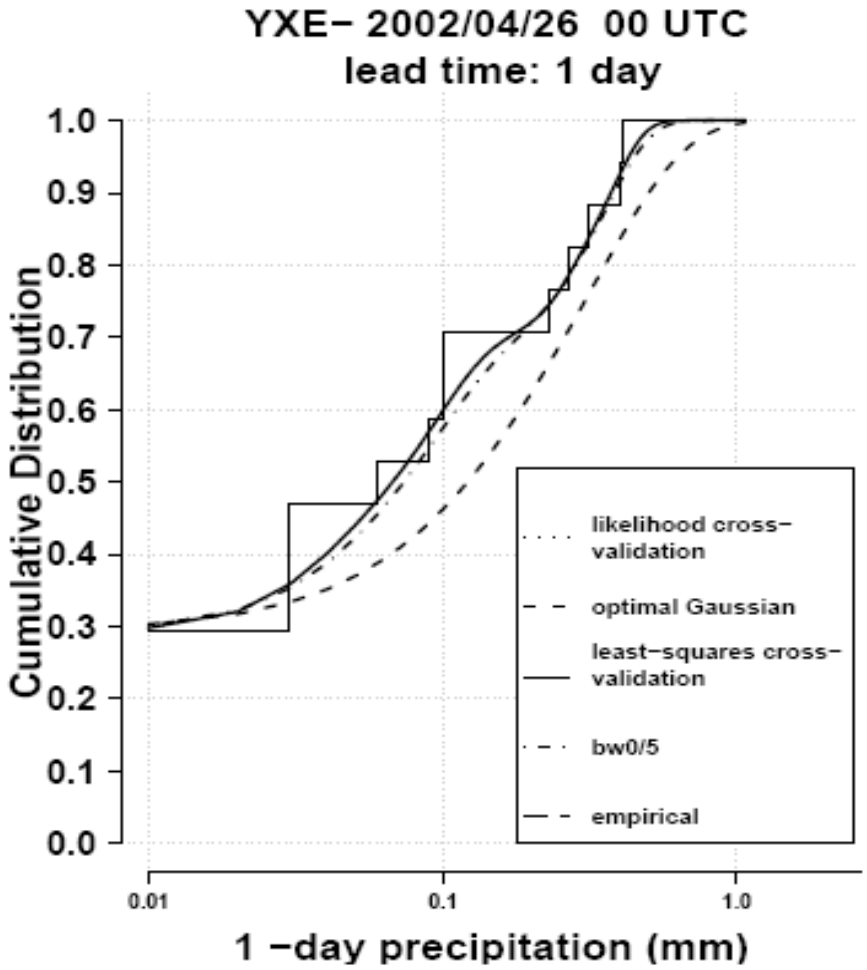


Continuous CDF

Advantages:

- For extreme event prediction, to estimate centile thresholds
- Assists with the ROC computation
- Simple to compare ensembles with different numbers of members

Many ways to construct a CDF →



From L. Wilson (EC)

Questions ?

- Provide examples where observations can be accurate but representativeness errors are large
- When is ensemble mean = ensemble average? Why it is used indistinctly for ensemble weather forecasts?
- Why does the ensemble mean forecast tend to have smaller RMSE than individual members?
- What are common approaches in numerical modeling to deal with uncertainties in the initial conditions? What are common approaches to deal with errors due to model limitations?

Outline

1. Background
   - Uncertainty in NWP systems
   - Ensemble Forecasting
2. Probabilistic forecast verification
   - Attributes of forecast quality
   - Performance metrics
3. Post-processing ensembles

Types of forecast

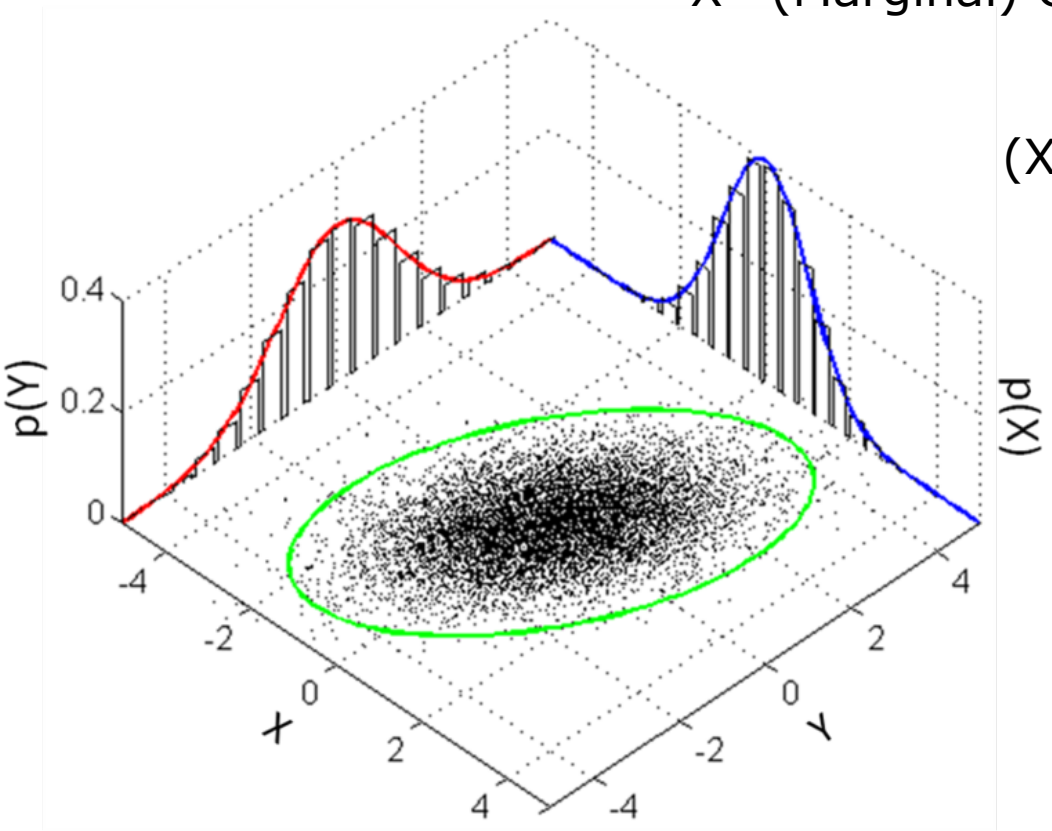
Categorical	Probabilistic
Yes/No	Assigns values between 0 and 100%
Verified with one single event	Requires many cases in which forecasts with X% probability are issued to be verified
Example: *Tomorrow's Max T will be 70F in College Park*.	Example: *There is a 30% chance of precipitation for tomorrow in College Park*

Probabilistic forecast verification

Comparison of a distribution of forecasts to a distribution of verifying observations

Y –(Marginal) Forecast distribution
X –(Marginal) Observation distribution

(X,Y) Joint distribution



Attributes of Forecast Quality

- **Accuracy:** Was the forecast close to what happened?
- **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 (e.g., chance, climatology, persistence)?

Performance measures

Brier Score

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)

Performance measures and Forecast attributions

Measures	Attributes
Brier Score, Brier Skill Score (BSS)	Accuracy
Reliability Diagram, Area of skill	Reliability, Resolution, Sharpness
Relative Operating Characteristic (ROC)	Discrimination
Rank Probability Score, Continuous Rank Probability Score (CRPS)	Integrated accuracy over the full PDF
Rank (Talagrand) Histograms	Spread assessment (outliers, biases)

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 Forecast accuracy of Binary Events.
- Range: 0 to 1. Perfect=0
- Weighs 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*:

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

reliability

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

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

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 Brier Skill Score

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

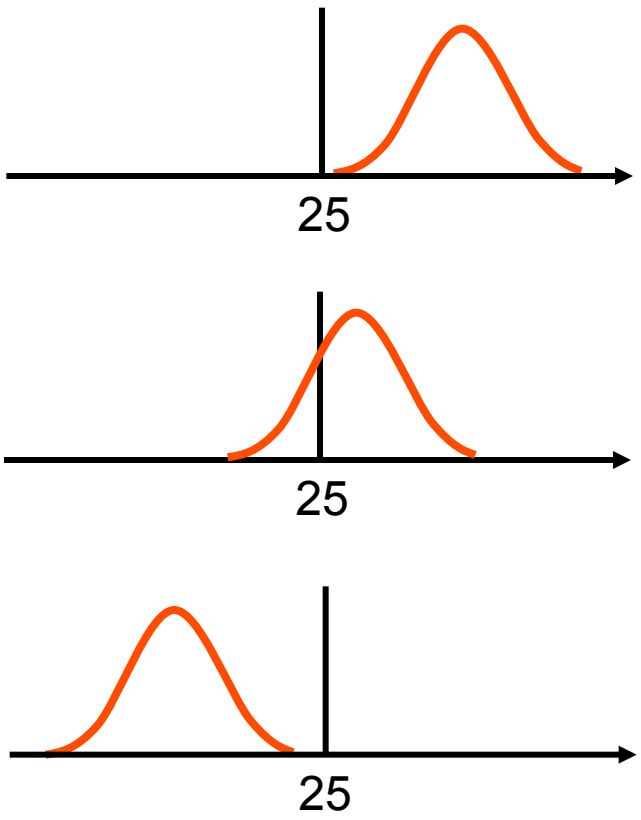
Reliability

- 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?

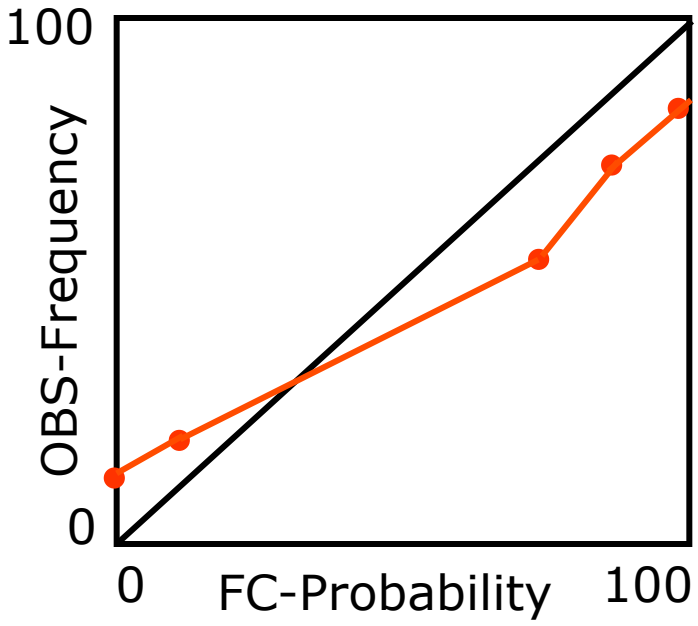


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

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?

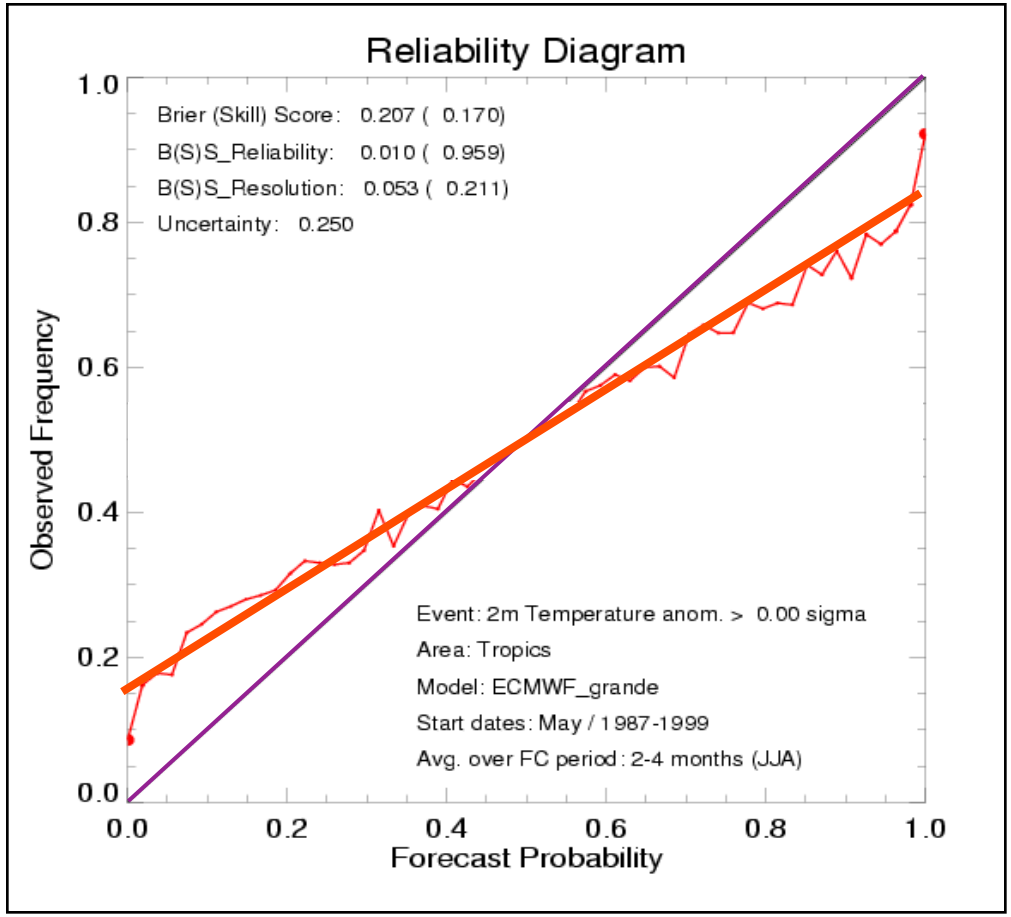


FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

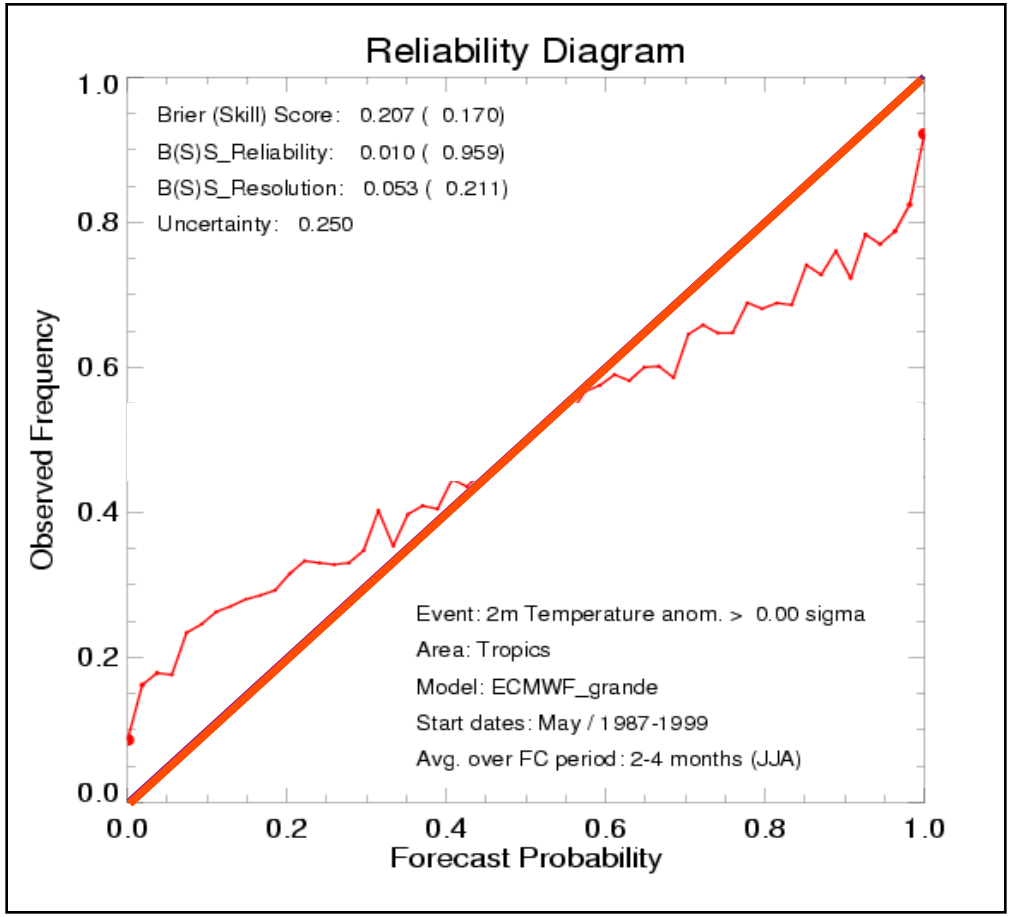
R. Hagedorn, 2007

Reliability Diagram

over-confident model



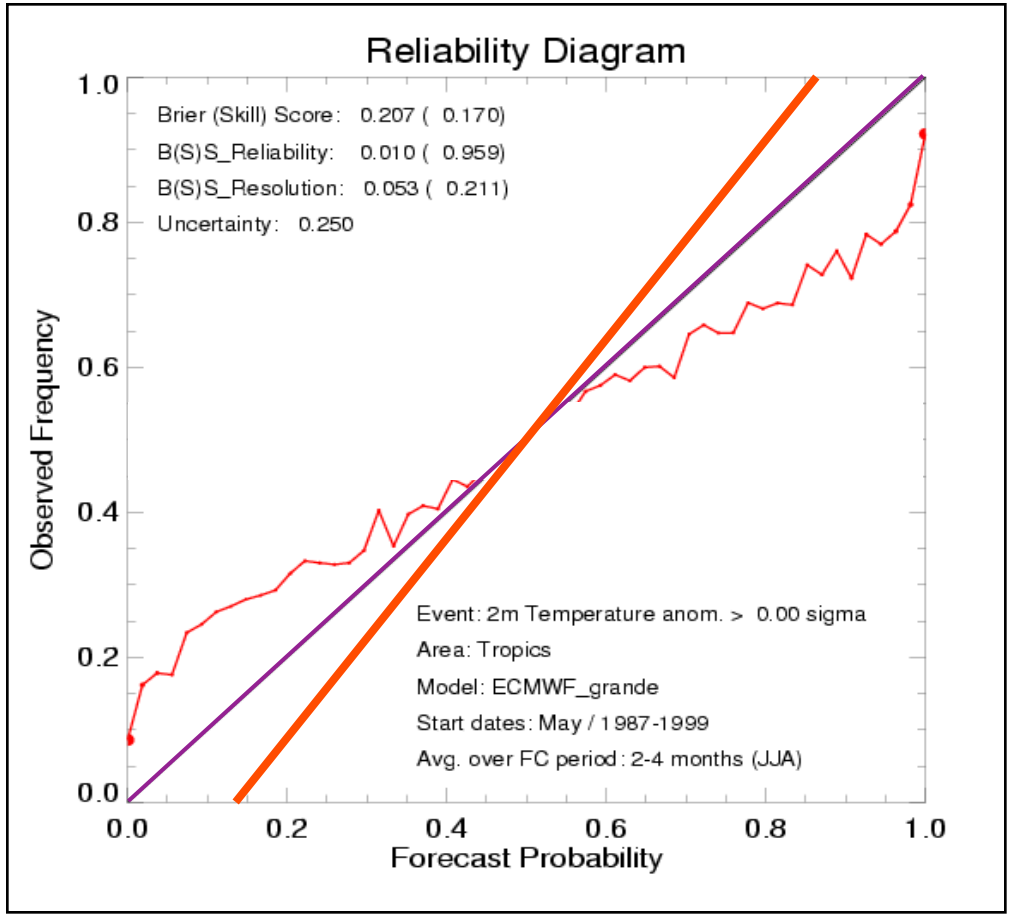
perfect model



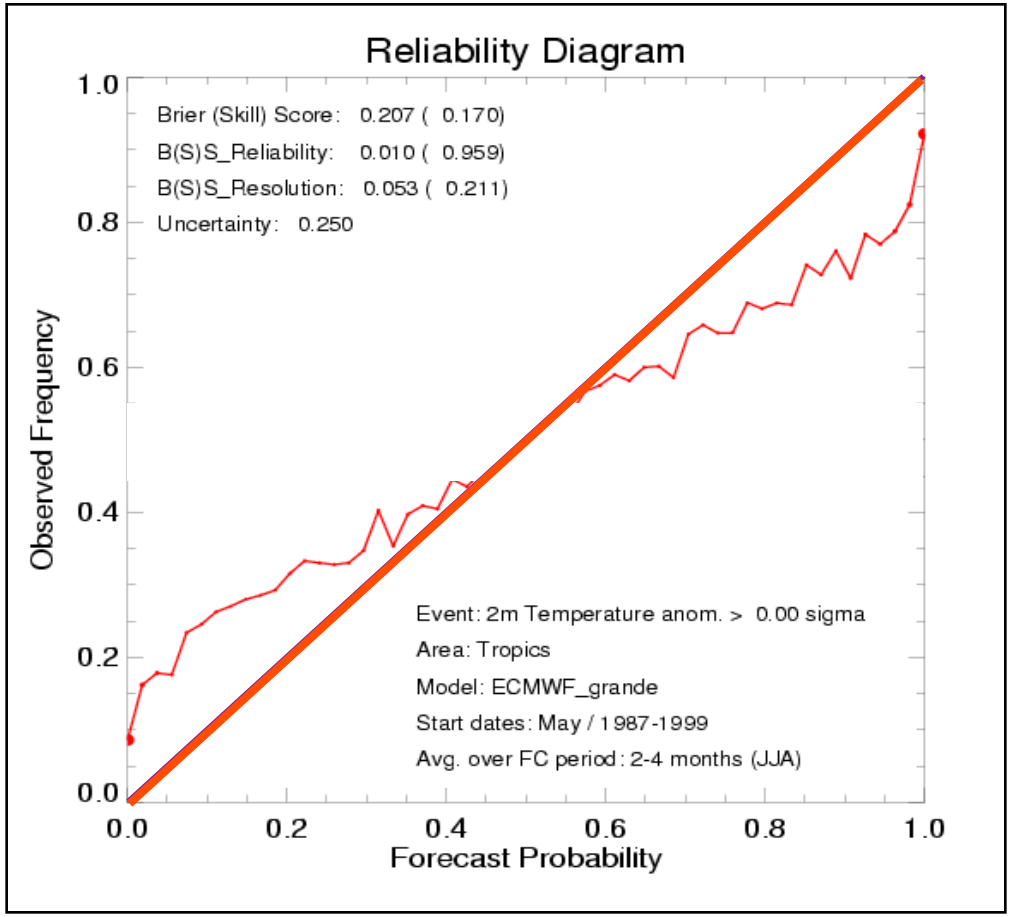
R. Hagedorn, 2007

Reliability Diagram

under-confident model



perfect model

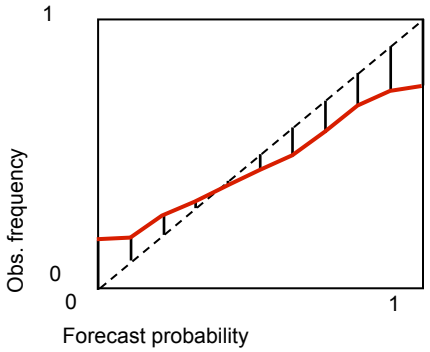


R. Hagedorn, 2007

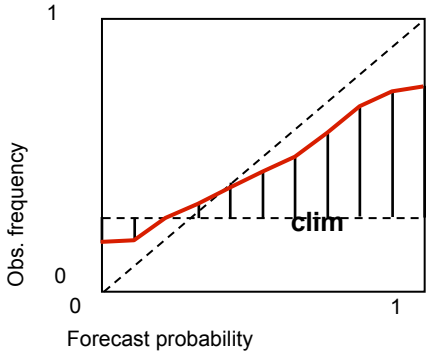
Reliability Diagram

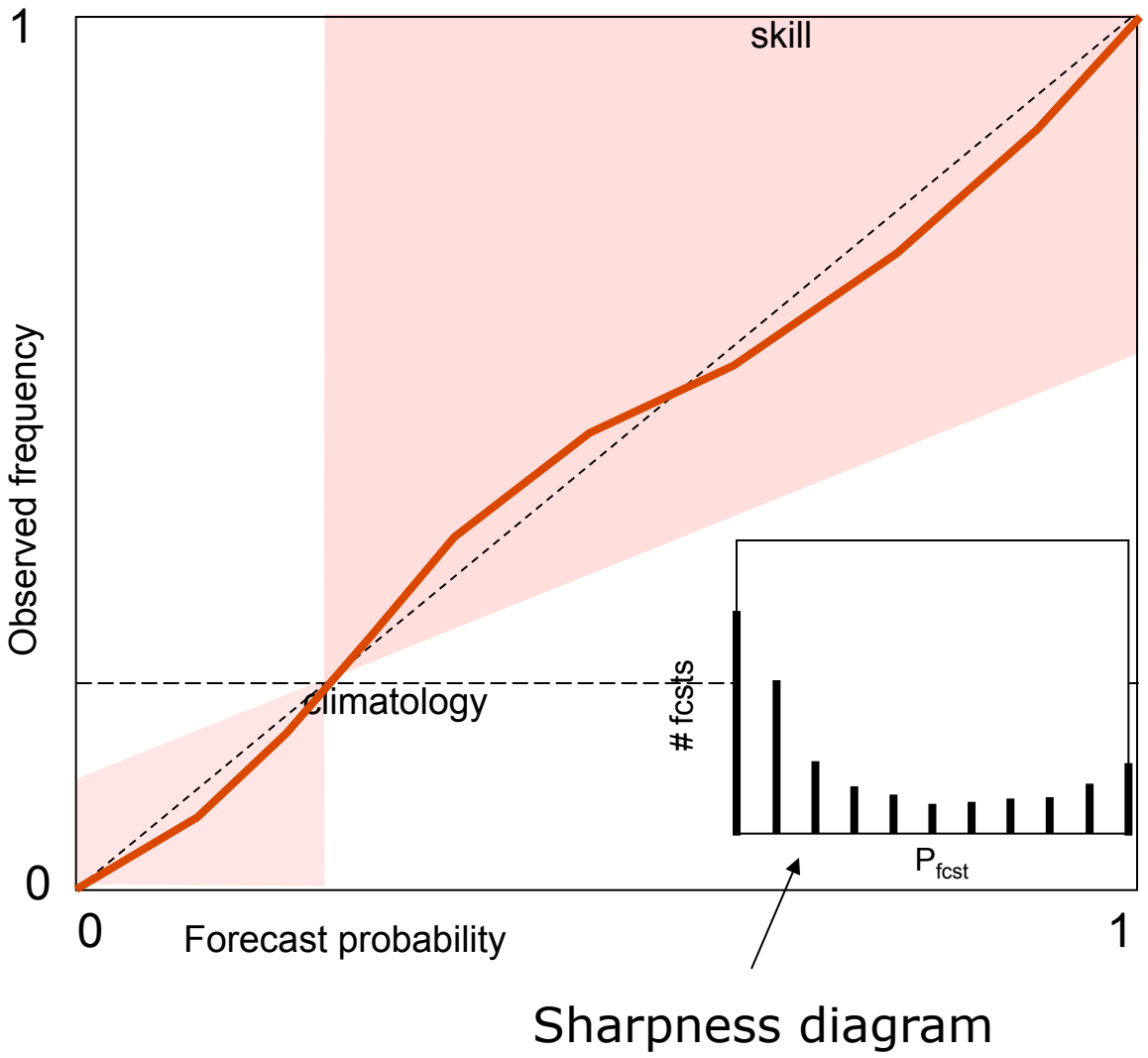
Reliability: How close to diagonal

(the lower the value the better)



Resolution: How far to horizontal (climatology) line

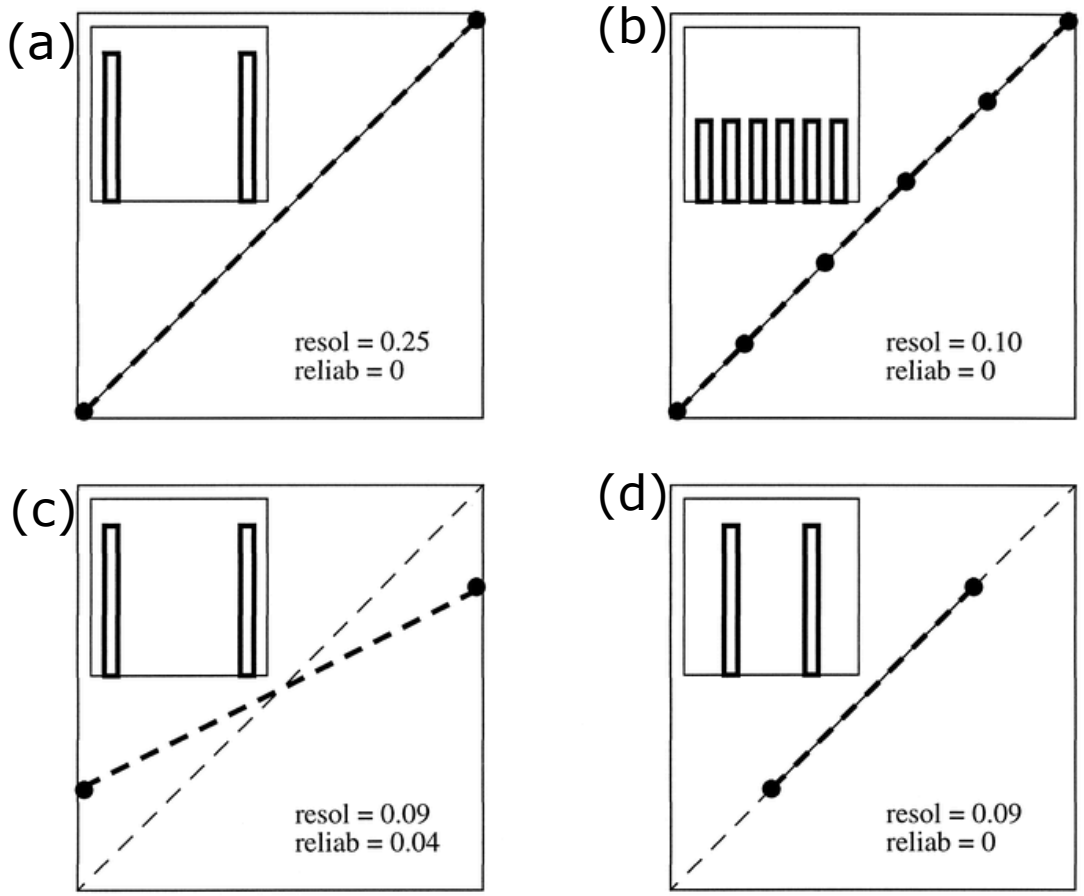




Sharpness diagram

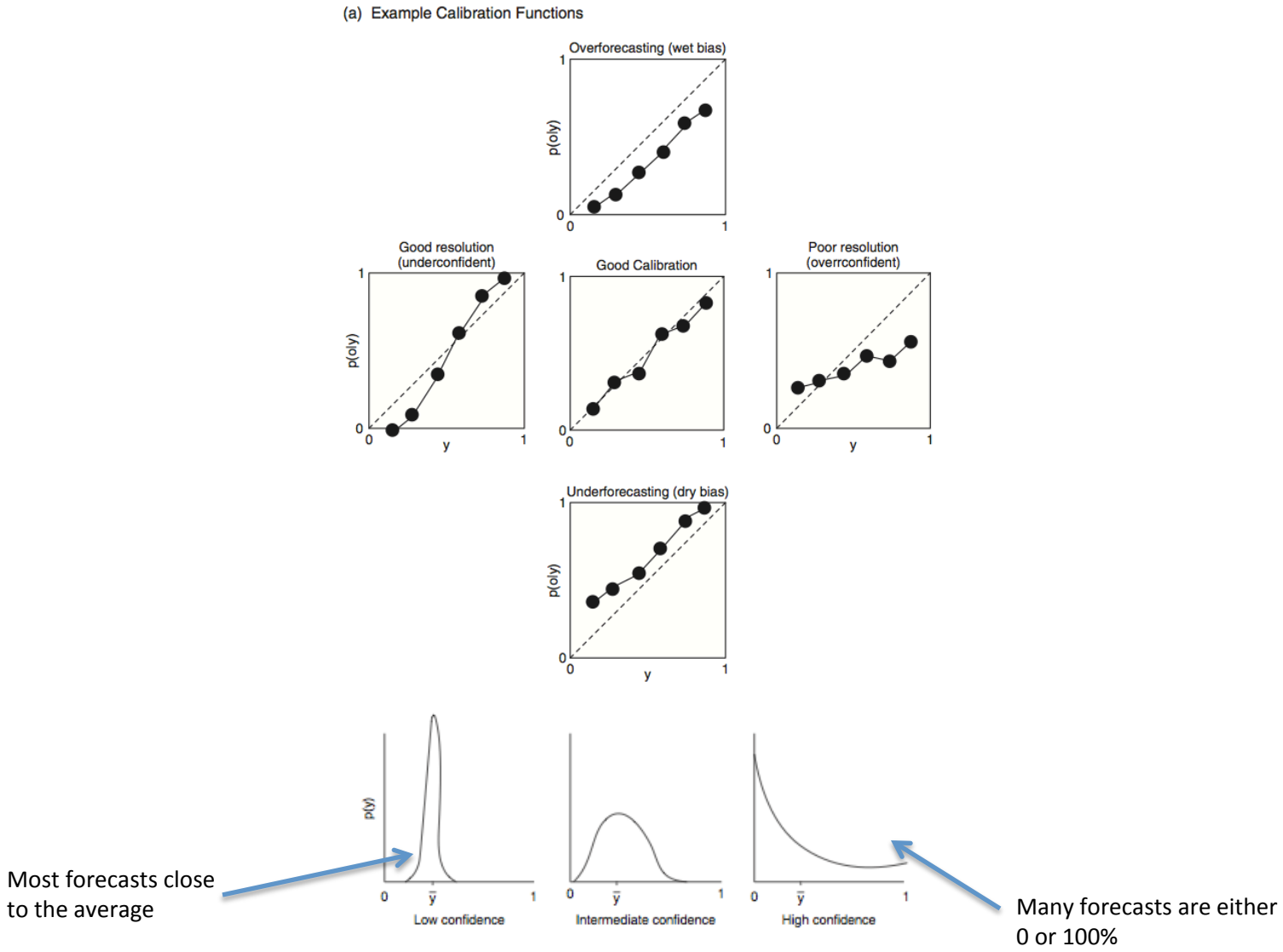
Examples of Reliability Diagram

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.

Examples of Reliability Diagram



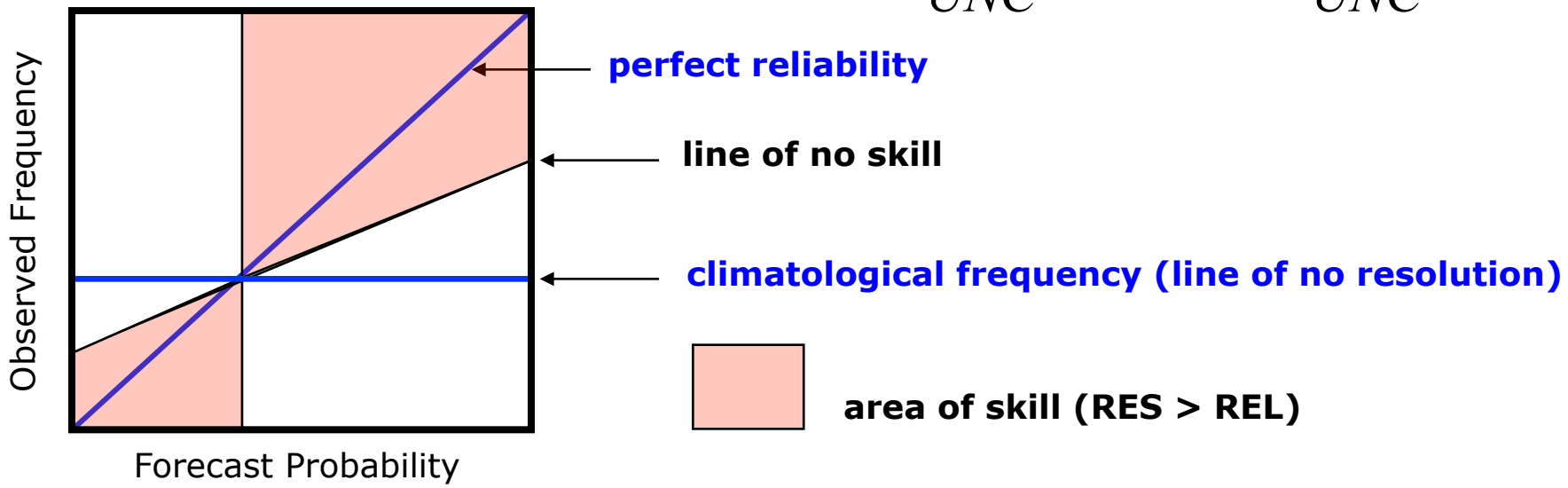
Wilks (2006)

Brier Skill Score & Reliability Diagram

- How to construct the area of positive skill?

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

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



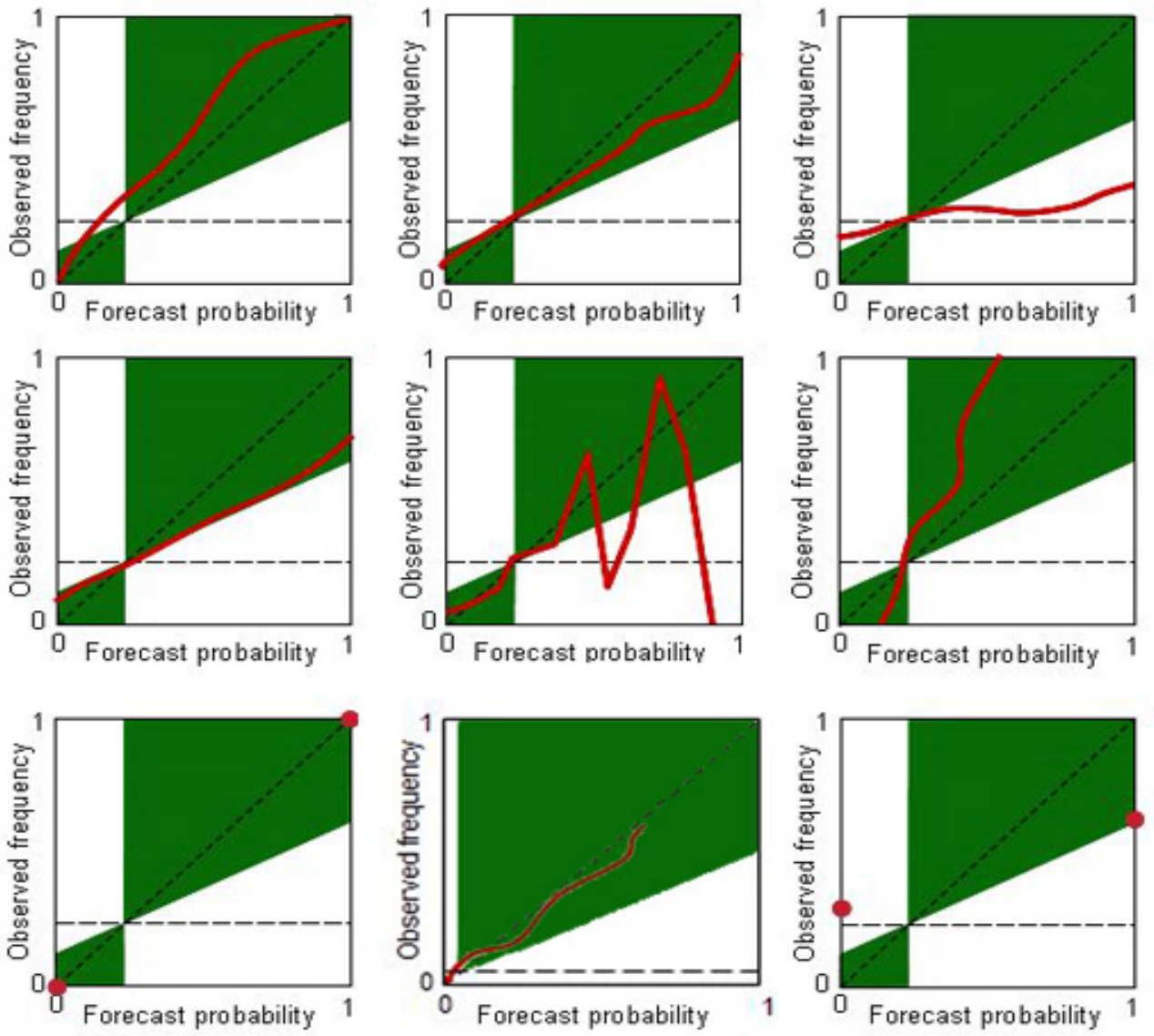
Reliability diagram: Construction

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*
   - $\textit{obs. relative frequency}_k = \textit{obs. occurrences}_k \;/\; \textit{num. forecasts}_k$
4. Plot observed frequency vs forecast probability
5. Plot sample climatology ("no resolution" line) (The sample base rate)
   - $\textit{sample climatology} = \textit{obs. occurrences} \;/\; \textit{num. forecasts}$
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 Exercise

Identify diagram(s) with:

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



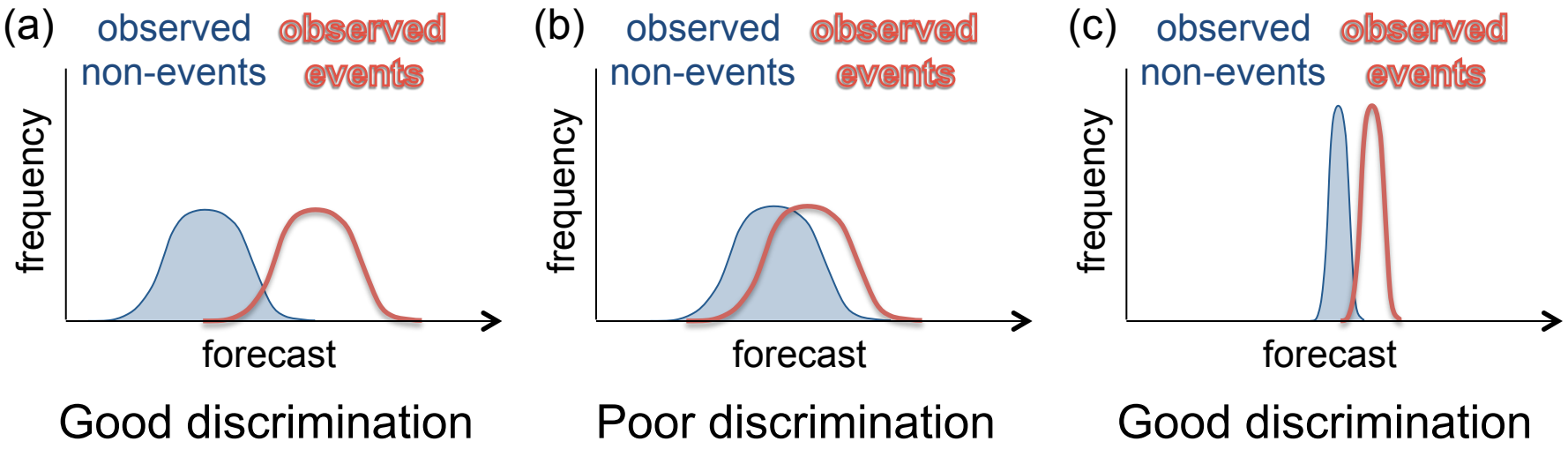
From L. Wilson (EC)

Reliability Diagrams - Comments

- Graphical representation of Brier score components
- Measures “reliability”, “resolution” and “sharpness”
- Sometimes called “attributes” diagram.
- Large sample size required to partition (bin) into subsamples conditional on forecast probability

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



From L. Wilson (EC)

Contingency Table

Suppose we partition the joint distribution (forecast , observation) according to whether an event occurred or not and ask the question whether it was predicted or not.

Forecast \ Observed	yes	no
yes	hits	false alarms
no	misses	Correct negatives

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

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

False Alarm Rate

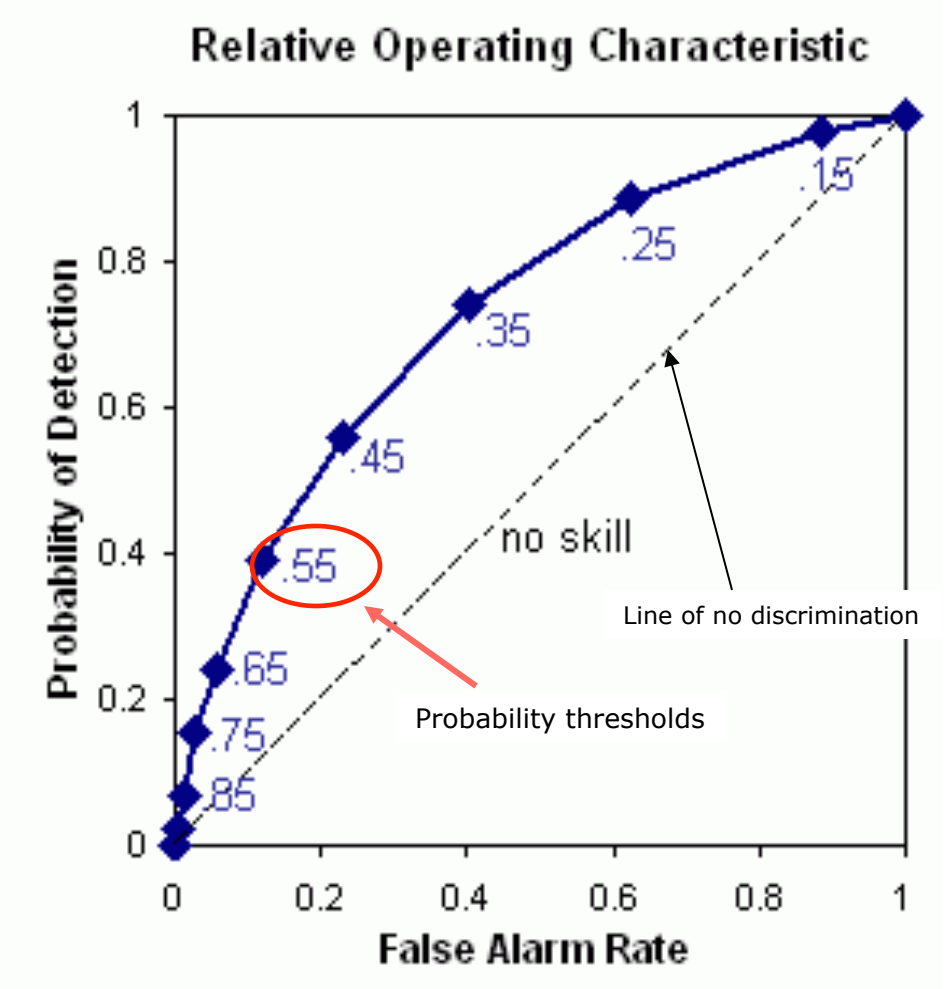
$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

Hit Rate

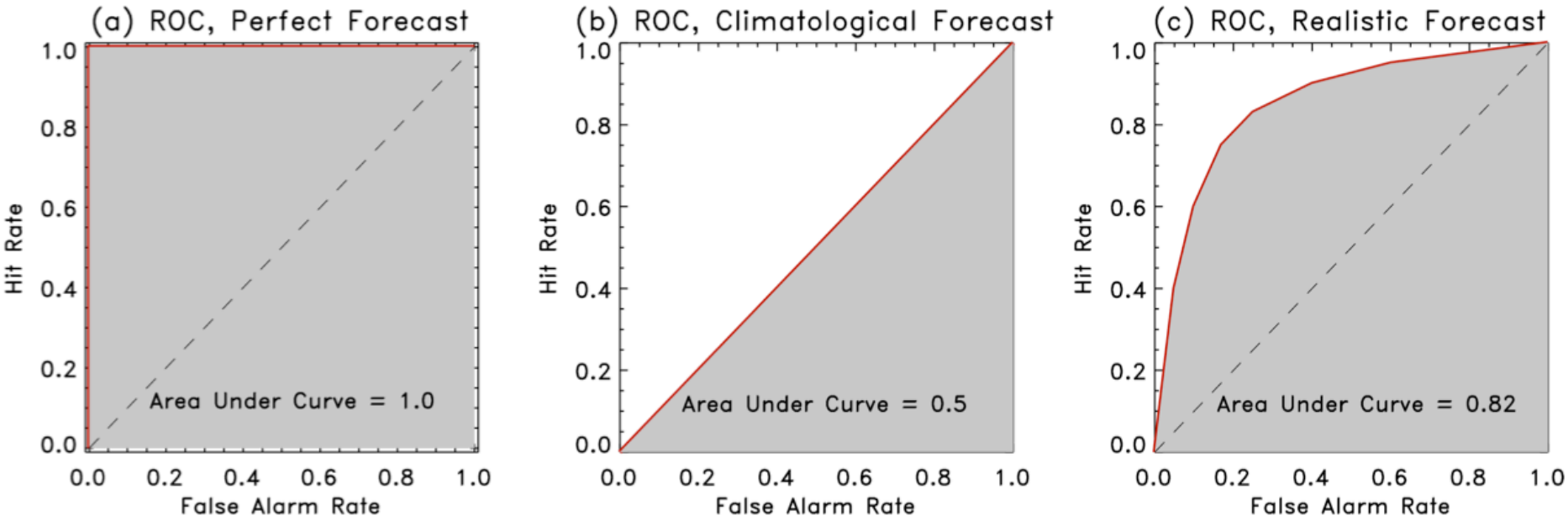
$$HR = \frac{hits}{hits + misses}$$

Construction of ROC curve

- Determine bins
  - There must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
- For each probability threshold, determine HR and FAR
- Plot HR vs FAR to give empirical ROC.



ROC - Interpretation



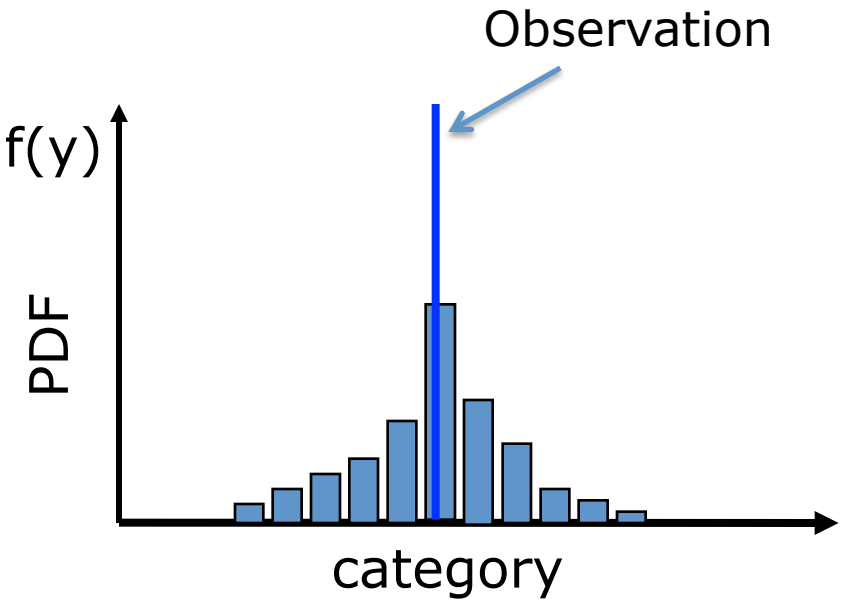
- Area under ROC curve (A) used as a single quantitative measure. Area range: 0 to 1. Perfect =1. No Skill = 0.5
- ROC Skill Score (ROCSS)

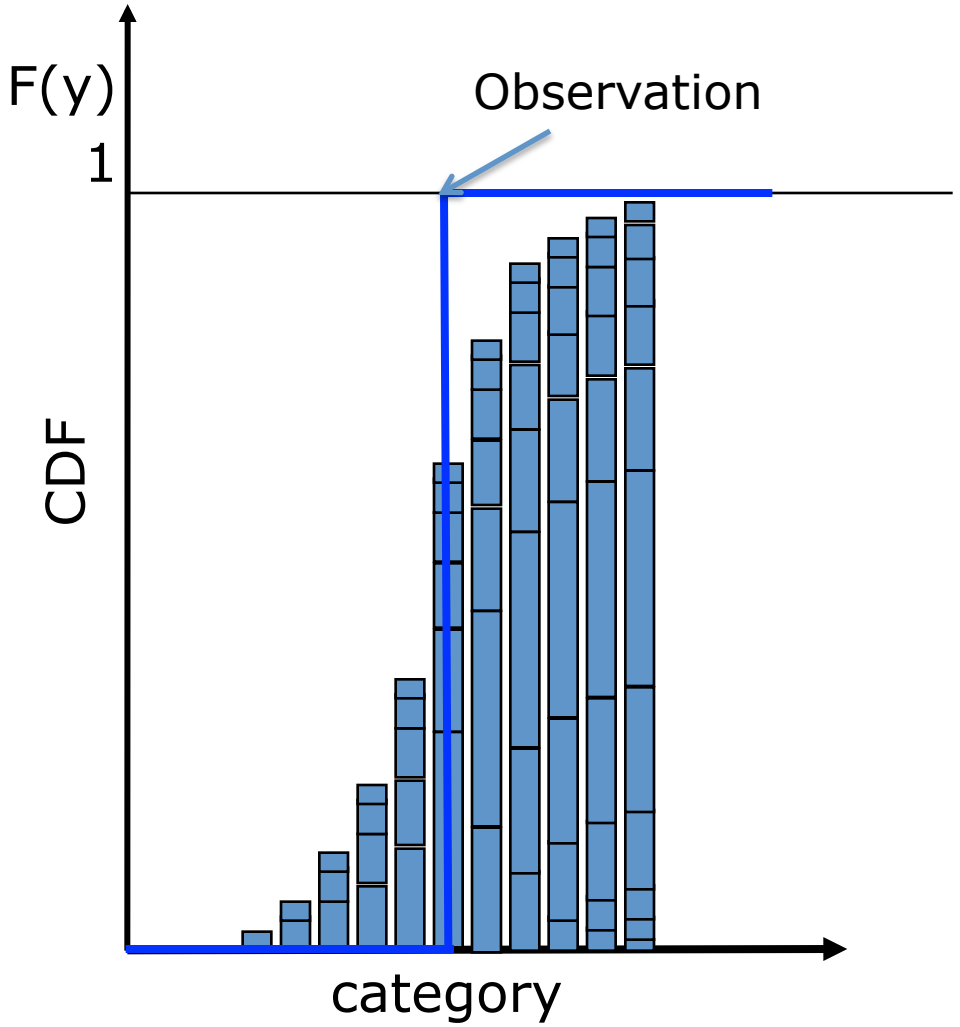
  ROCSS = 2A – 1

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
- Allows the performance comparison between probability and deterministic forecasts

Rank Probability Score

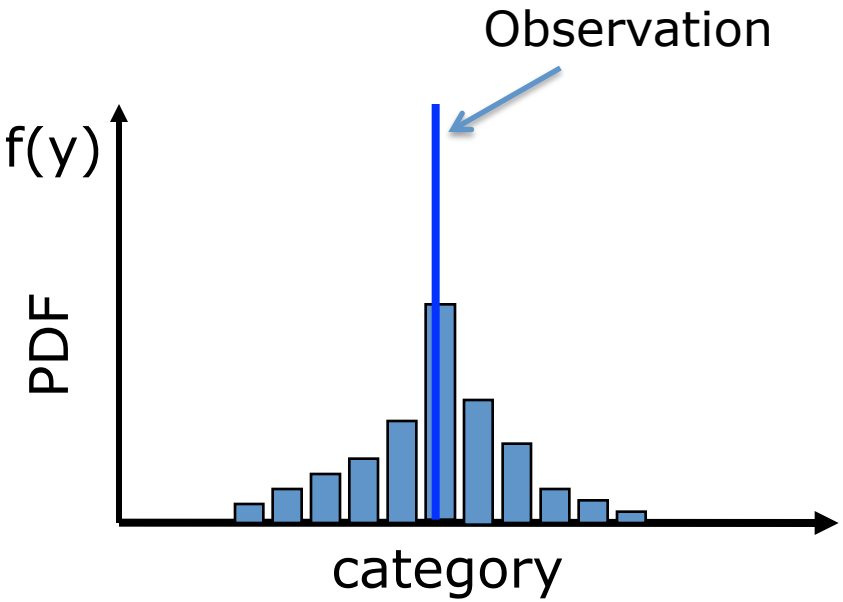




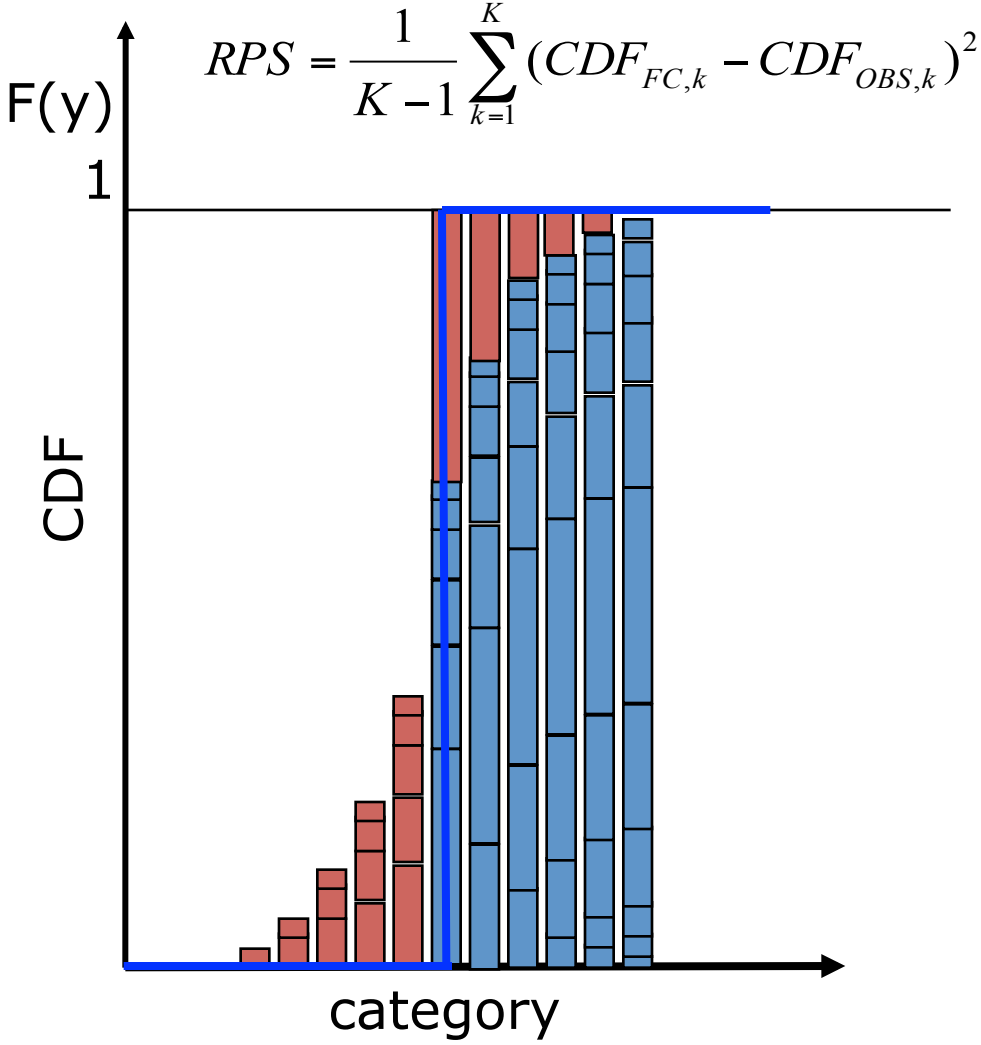
Measures the distance between the Observation and the Forecast probability distributions

R. Hagedorn, 2007

Rank Probability Score

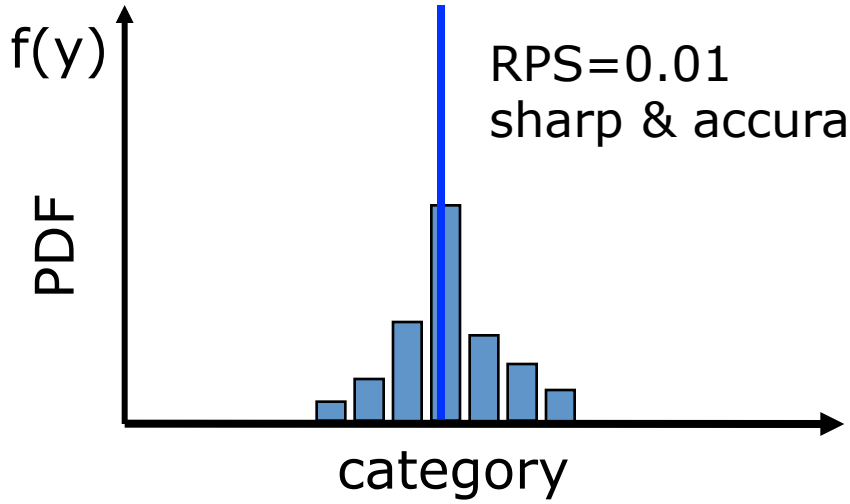


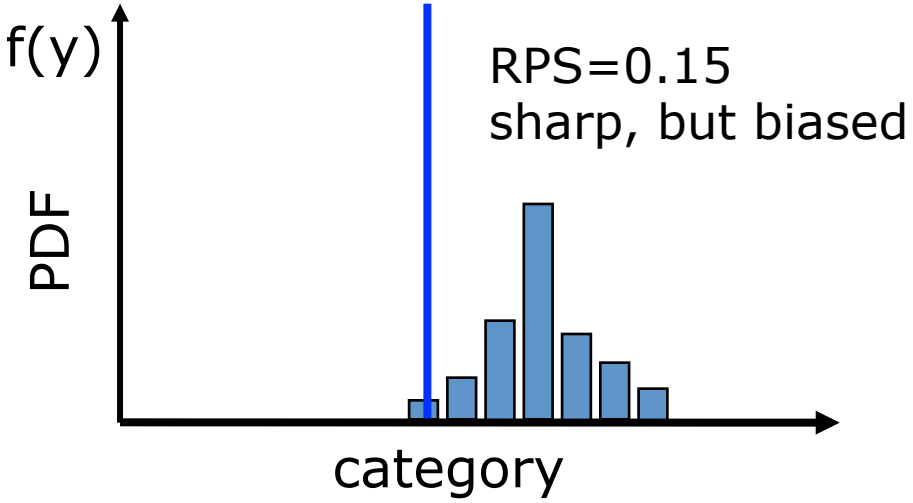
$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{FC,k} - CDF_{OBS,k})^2$$

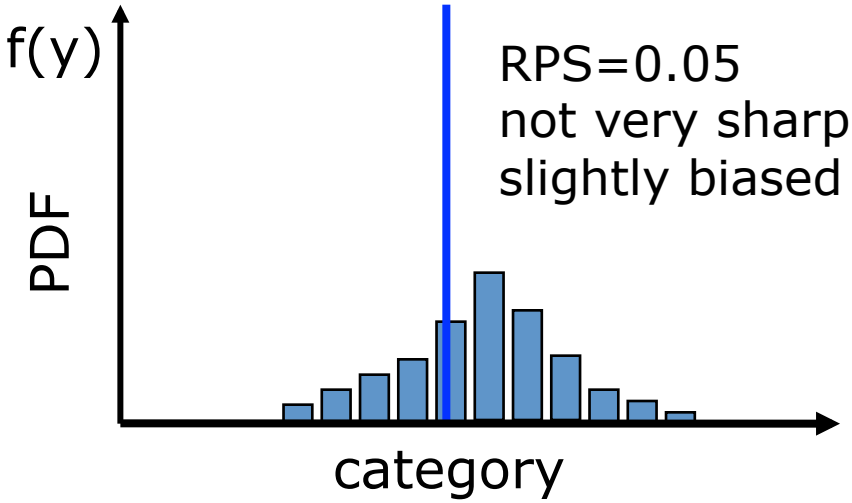


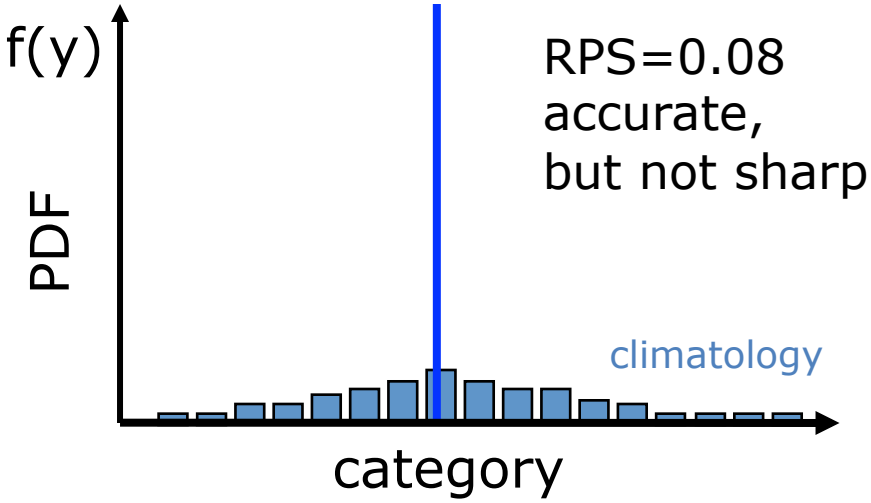
- Measures the quadratic distance between forecast and verification probabilities for several probability categories *k*. Range: 0 to 1. Perfect=0
- Emphasizes accuracy by penalizing large errors more than "near misses"
- Rewards sharp forecast if it is accurate

Rank Probability Score



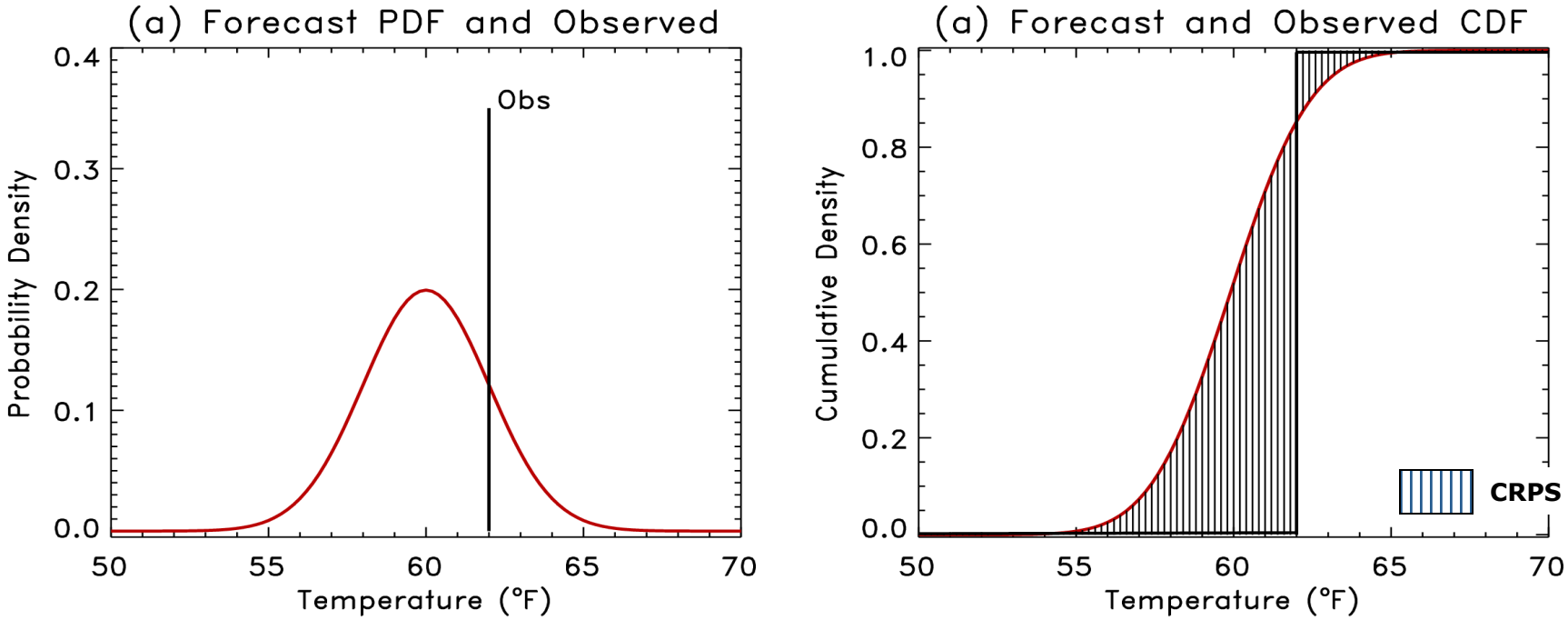






Continuous Rank Probability Score

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

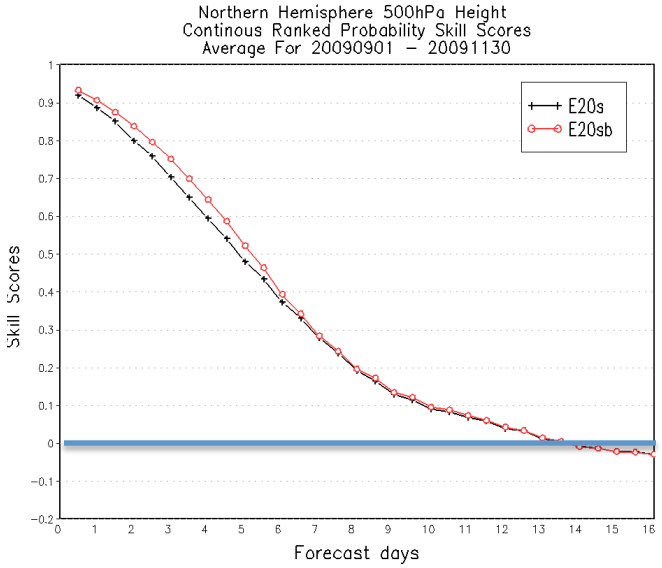


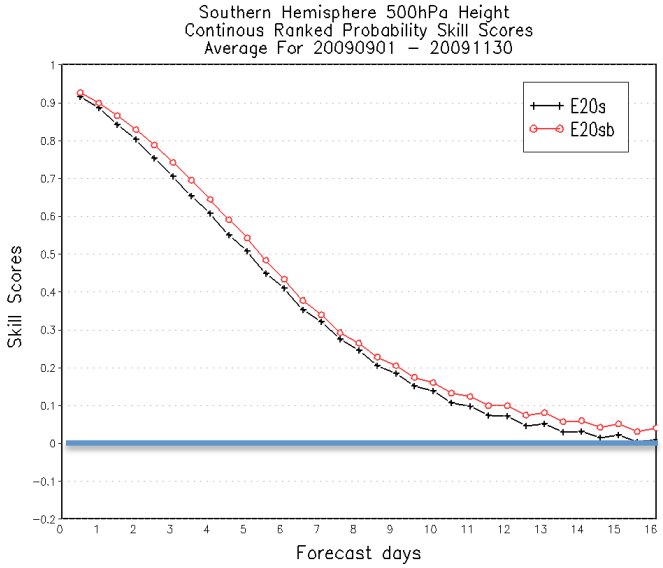
- Area difference between CDF observation and CDF forecast
- Defaults to MAE for deterministic forecast
- Flexible, can accommodate uncertain observations

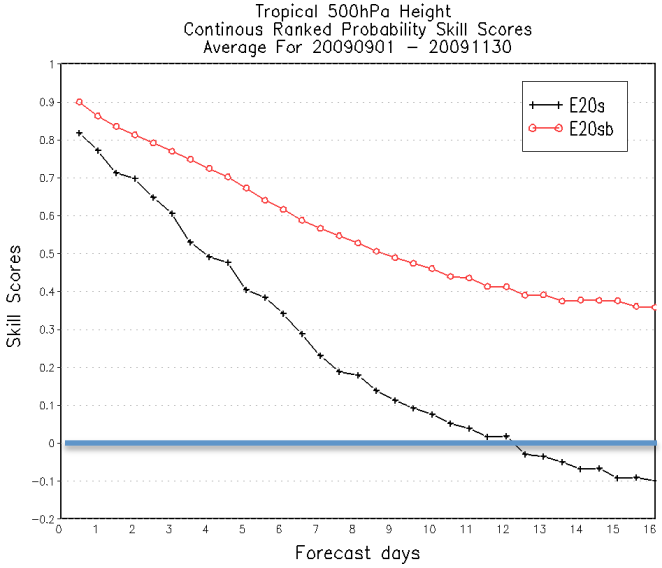
Continuous Rank Probability Skill Score

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

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



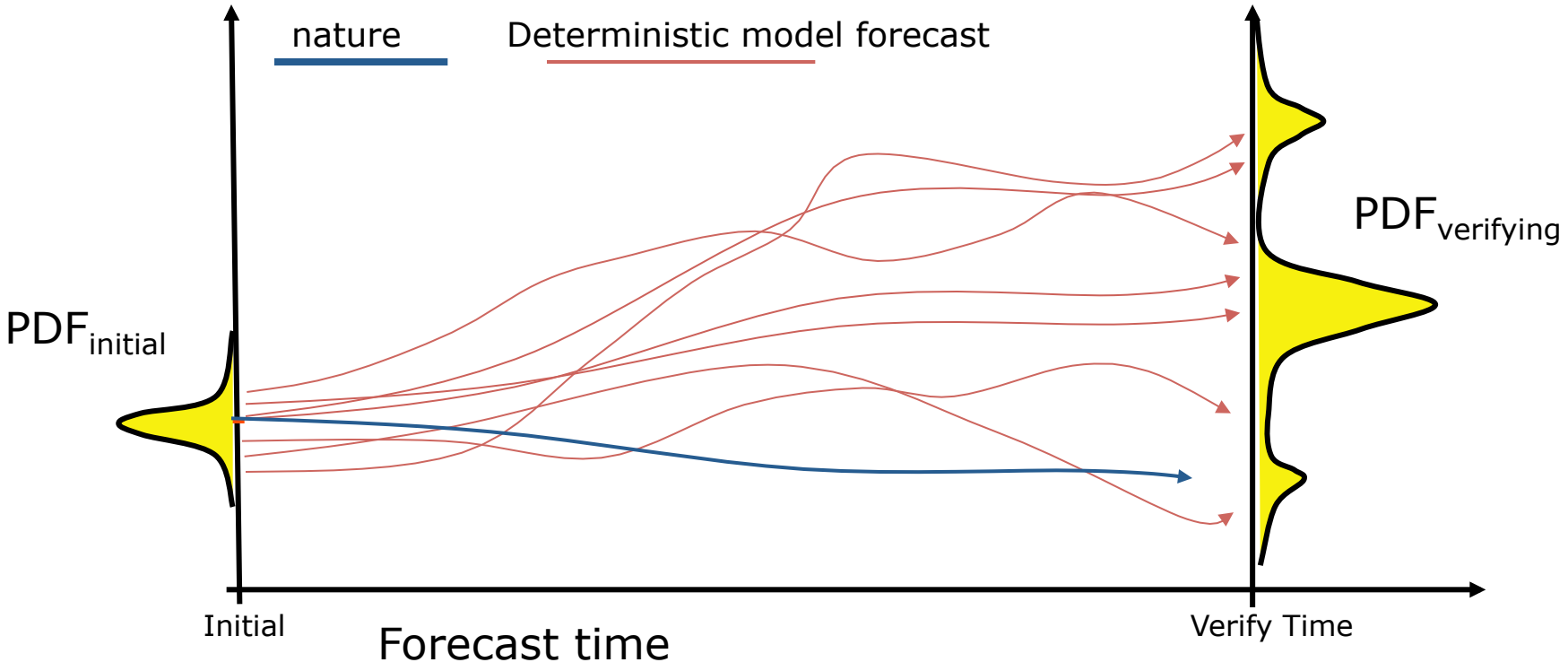




Courtesy of Y. Zhu (EMC/NCEP)

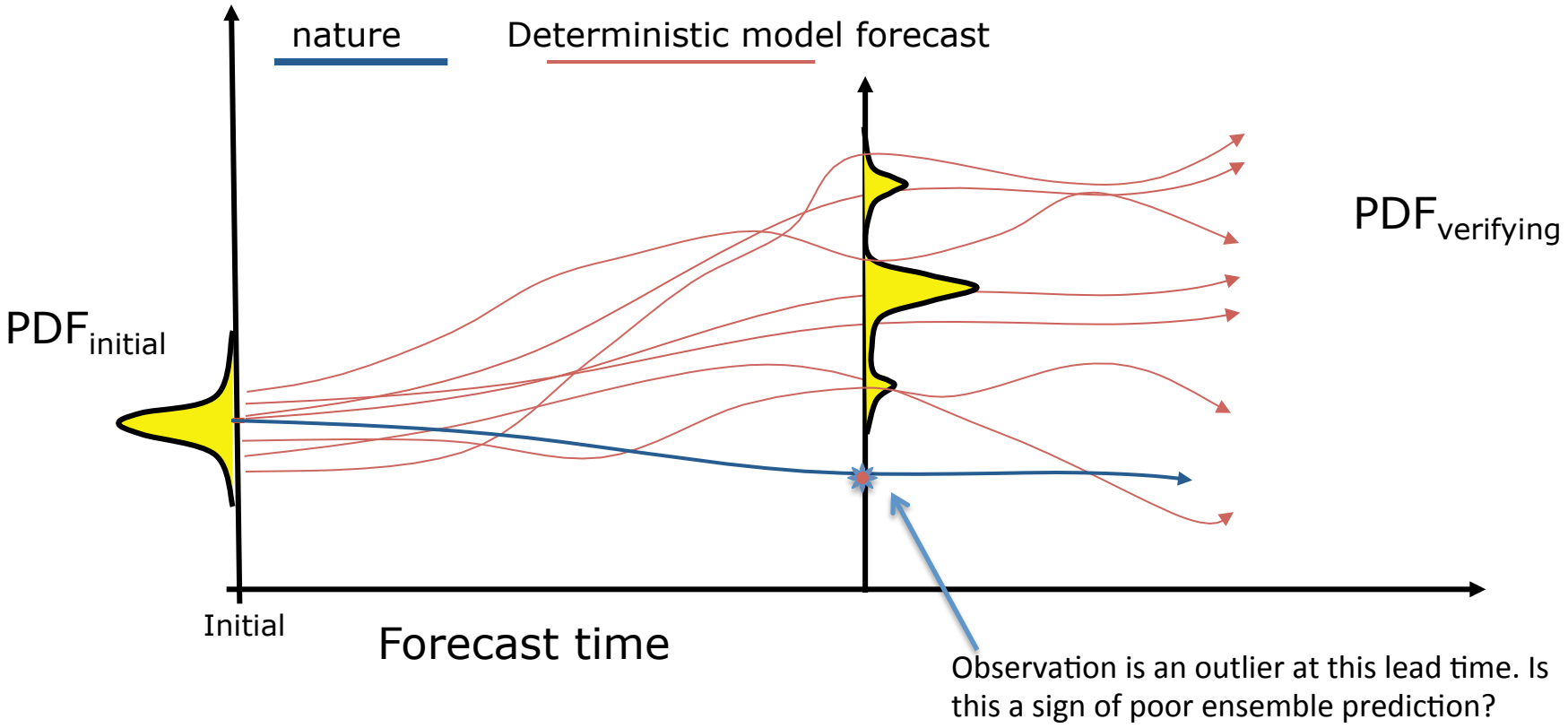
Rank Histogram (aka Talagrand Diagram)

Do the observations statistically belong to the distributions of the forecast ensembles?



Rank Histogram (aka Talagrand Diagram)

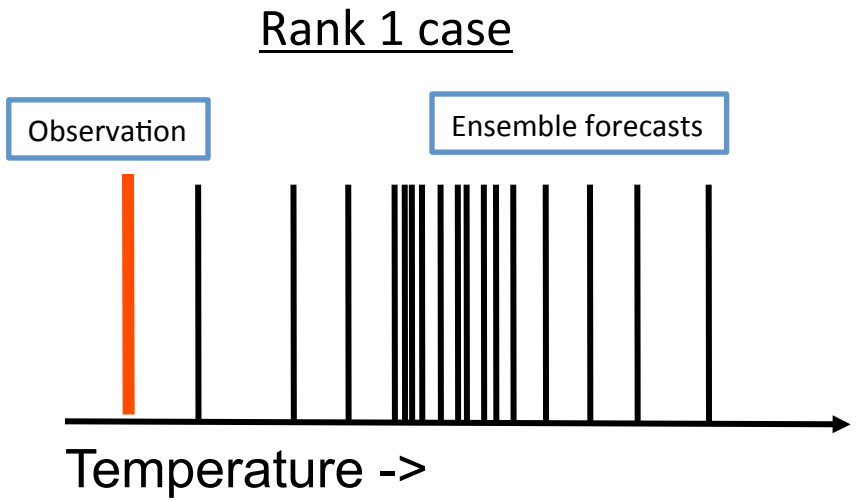
Do the observations statistically belong to the distributions of the forecast ensembles?



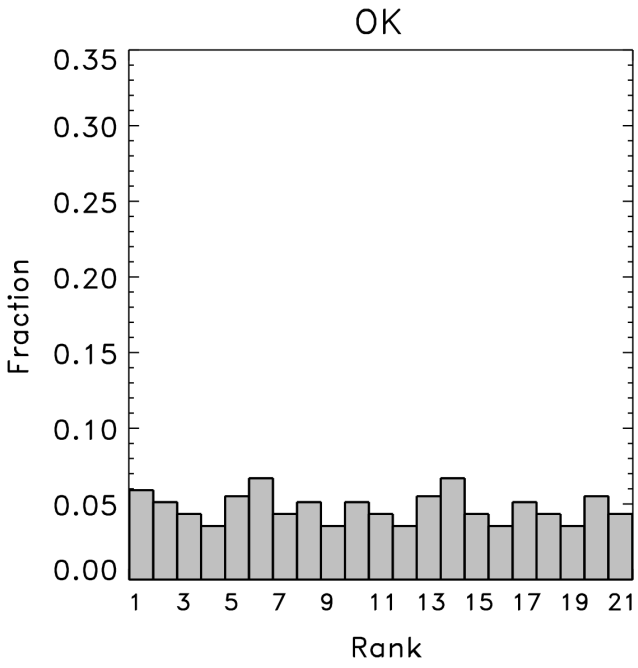
Rank Histogram

Rank Histograms asses whether the ensemble spread is consistent with the assumption that the observations are statistically just another member of the forecast distribution.

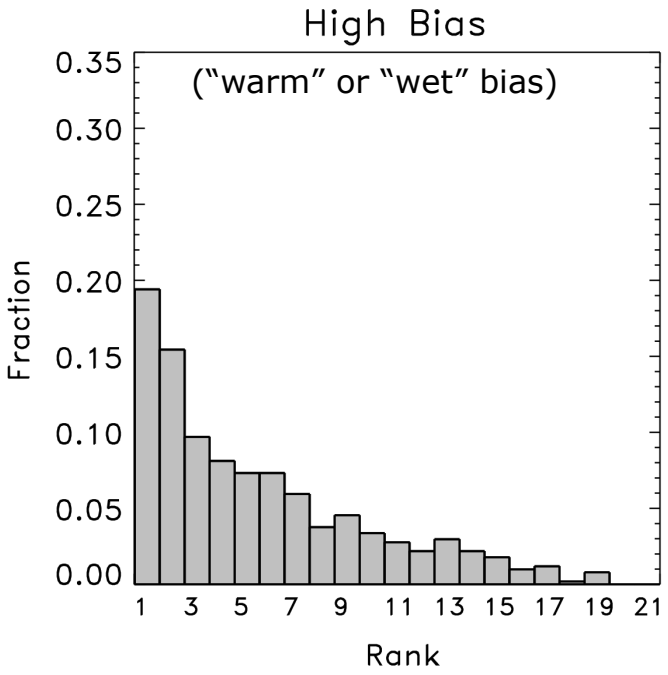
Procedure: Sort ensemble members in increasing order and determine where the verifying observation lies with respect to the ensemble members



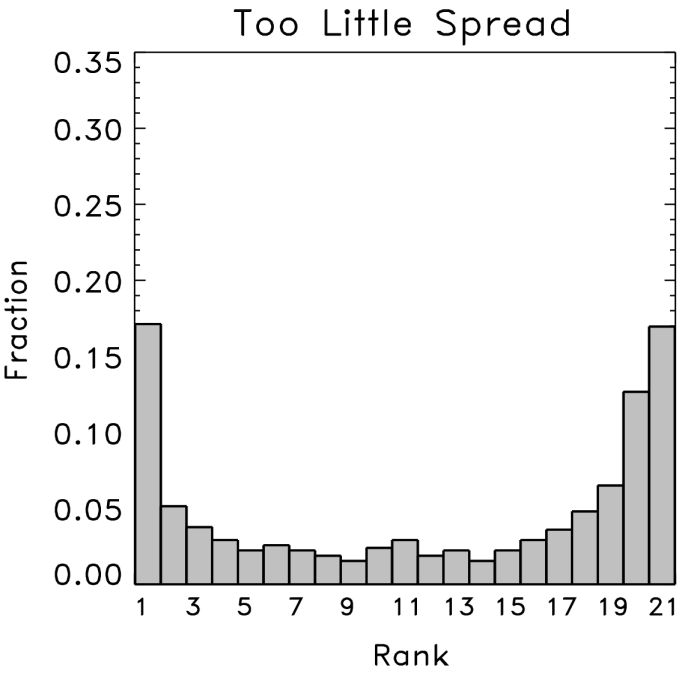
Rank Histograms



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 S_k is the number of occurrences in the kth interval of the histogram.

Outline

1. Background
   - Uncertainty in NWP systems
   - Introduction to Ensemble Forecasts
2. Probabilistic forecast verification
   - Attributes of forecast quality
   - Performance metrics
3. Post-processing ensembles
   - Distribution fitters
   - PDF Calibration
   - Combination

Estimating Probabilities from ensembles

- We want to extract as much information as possible from the set of discrete values generated by the ensemble.
- Approaches to convert discrete values to PDF or CDF:
  - Discrete
  - Histograms
  - Continuous functions
    - parametric
    - nonparametric

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 Ensemble Prediction systems

- 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

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

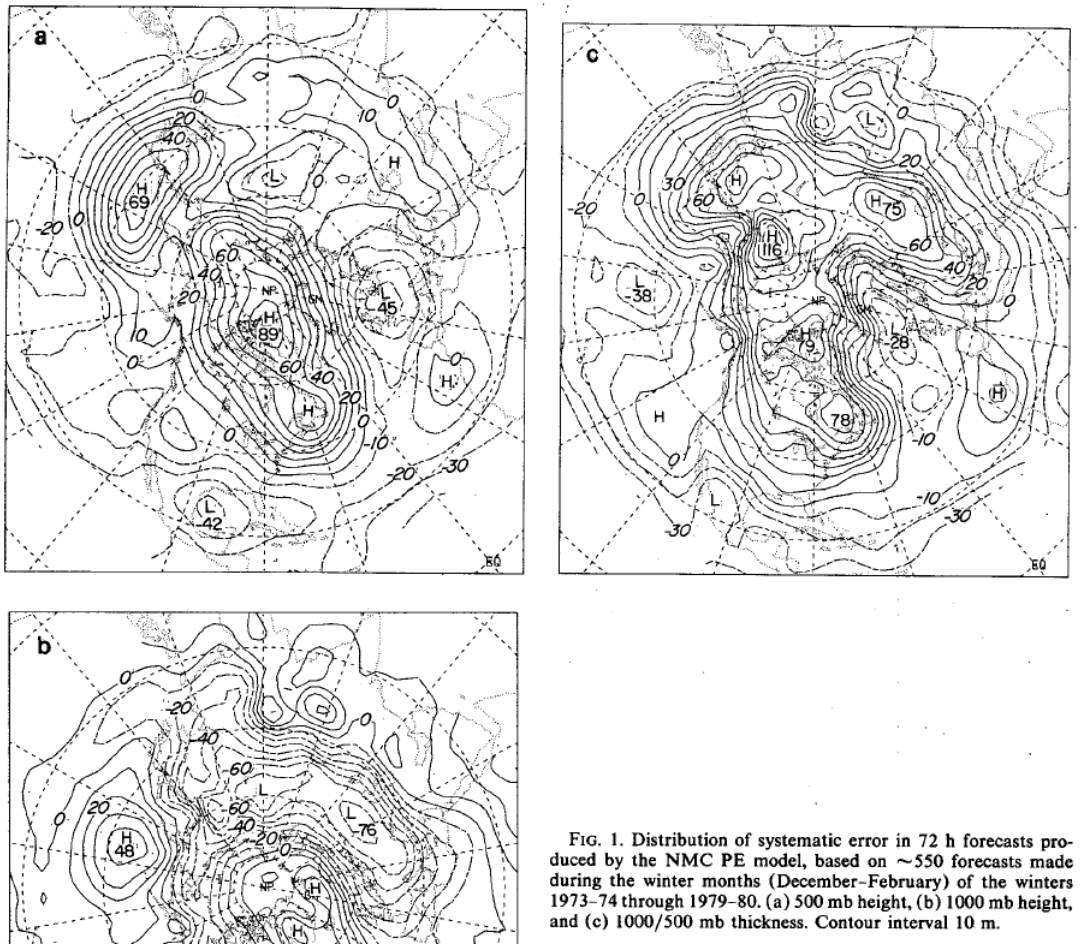
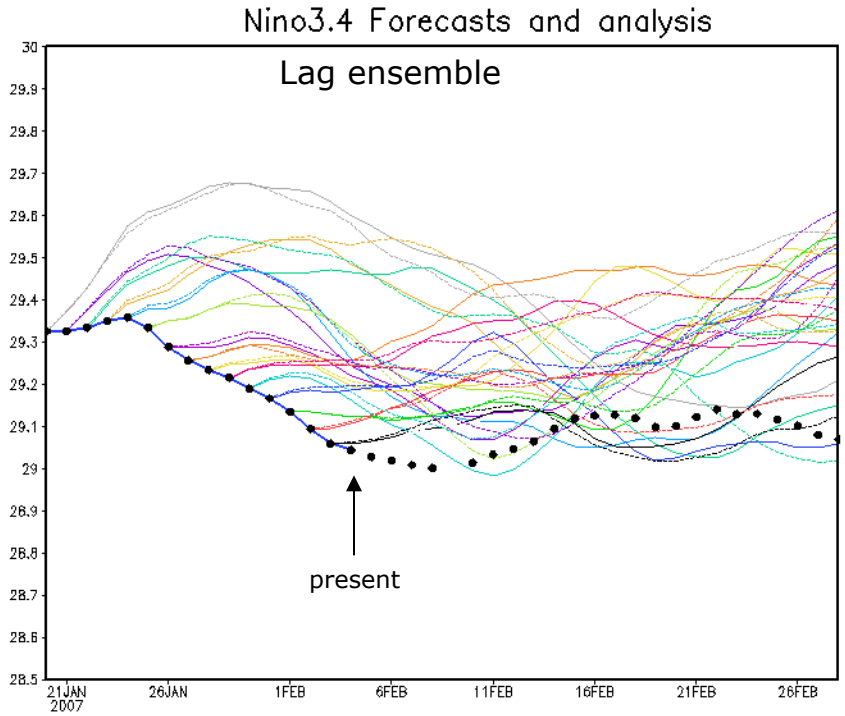


FIG. 1. Distribution of systematic error in 72 h forecasts produced by the NMC PE model, based on ~550 forecasts made during the winter months (December–February) of the winters 1973–74 through 1979–80. (a) 500 mb height, (b) 1000 mb height, and (c) 1000/500 mb thickness. Contour interval 10 m.



Note "Warm" tendency of the model

Bias correction

A first order bias correction approach is by computing the mean error from historical forecast-observation pairs archive:

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

with: $\boldsymbol{e}_i$ = ensemble mean of the ith forecast

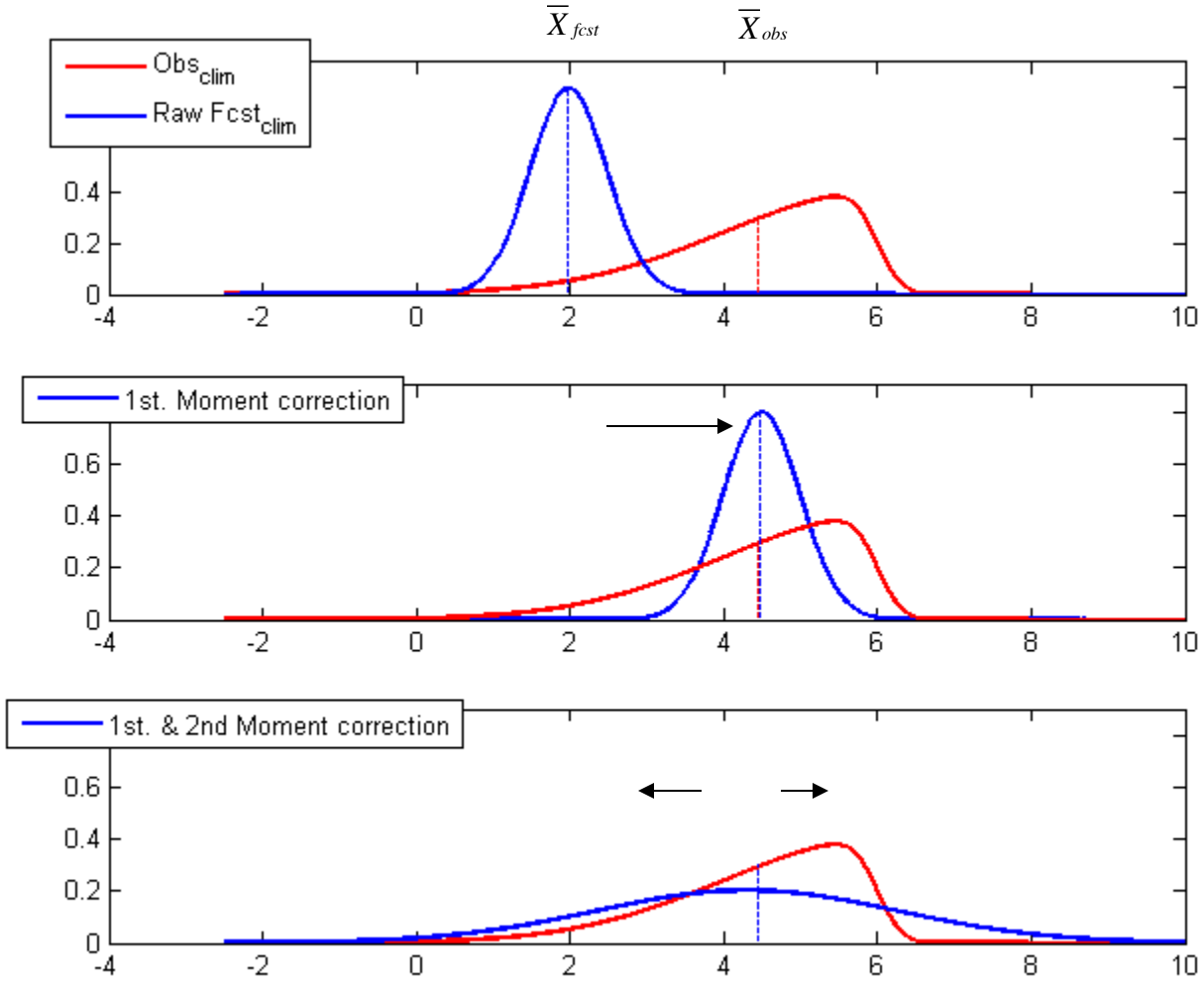
$\boldsymbol{o}_i$ = value of ith observation

N = number of observation-forecast pairs

• This systematic error is removed from each ensemble member forecast. Notice that the 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



$\overline{X}_{fcst} = \overline{X}_{obs}$

$\sigma_{fcst} = \sigma_{obs}$

Quantile Mapping Method

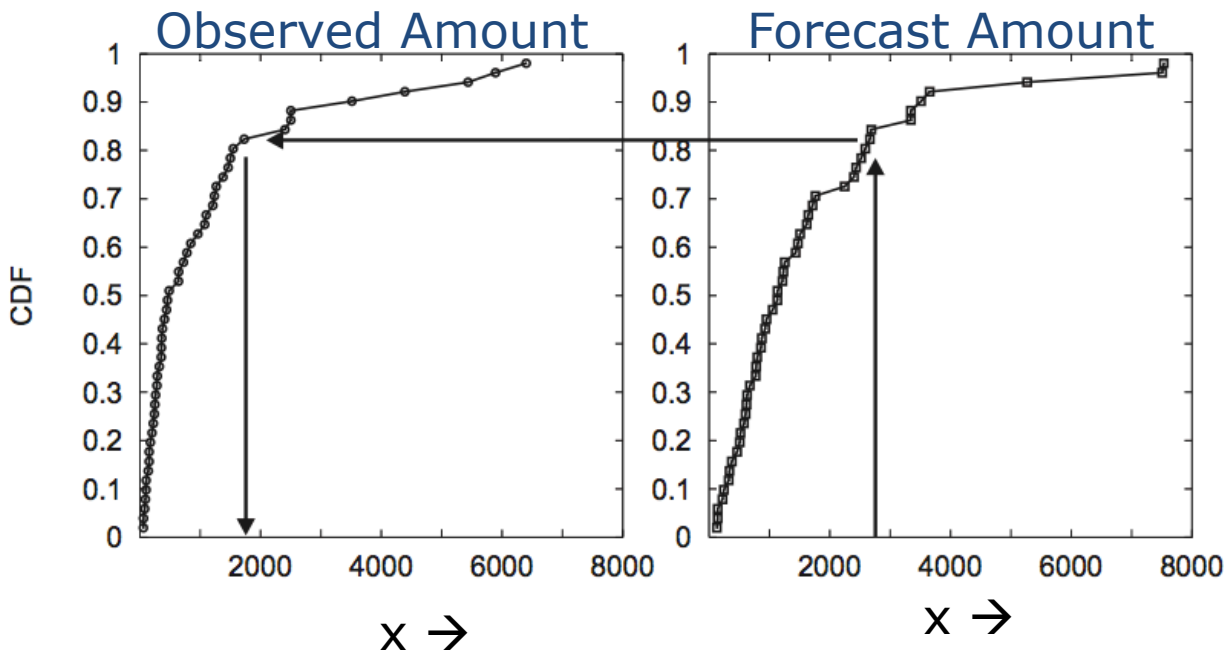
Uses the empirical Cumulative Distribution Function for observed and corresponding forecast to remove the bias

Let *G* be the cdf of the observed quantity. Let *F* be the cdf of the corresponding historical forecast.

The corrected ensemble forecast must satisfy:

$$Z = G^{-1}(F(x))$$

x is the forecast amount.



In this example, a forecast amount of 2800 is transformed into 1800 for the 80 percentile

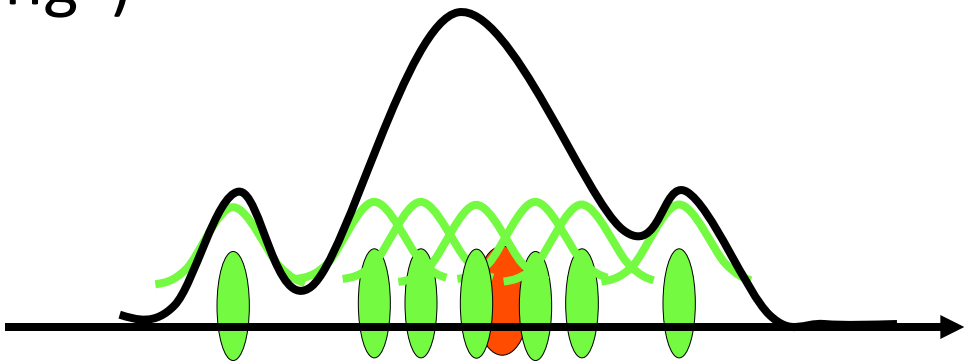
Multiple implementation of det. MOS

- 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

Definition of Kernel

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

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

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

- Examples:

Uniform

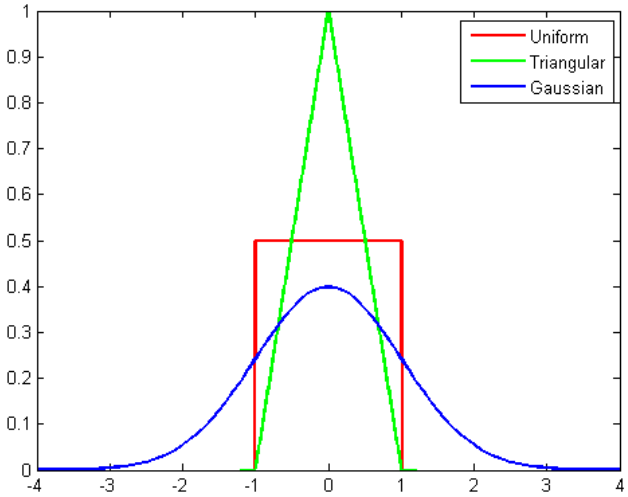
$$K(u) = \frac{1}{2}\mathbf{1}_{\{|u|\le 1\}}$$

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:

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

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.

Multi-Model Combination (Consolidation)

- Making the best single forecast out of a number of forecast inputs
- Necessary as large supply of forecasts available
- Expressed as a linear combination of participant models:

$$C = \sum_{k=1}^{K} \alpha_k \zeta_k$$

- K number of participating models
- ζ input forecast at a particular initial month and lead time

Task: Finding *K* optimal weights, α_k , corresponding to each input model

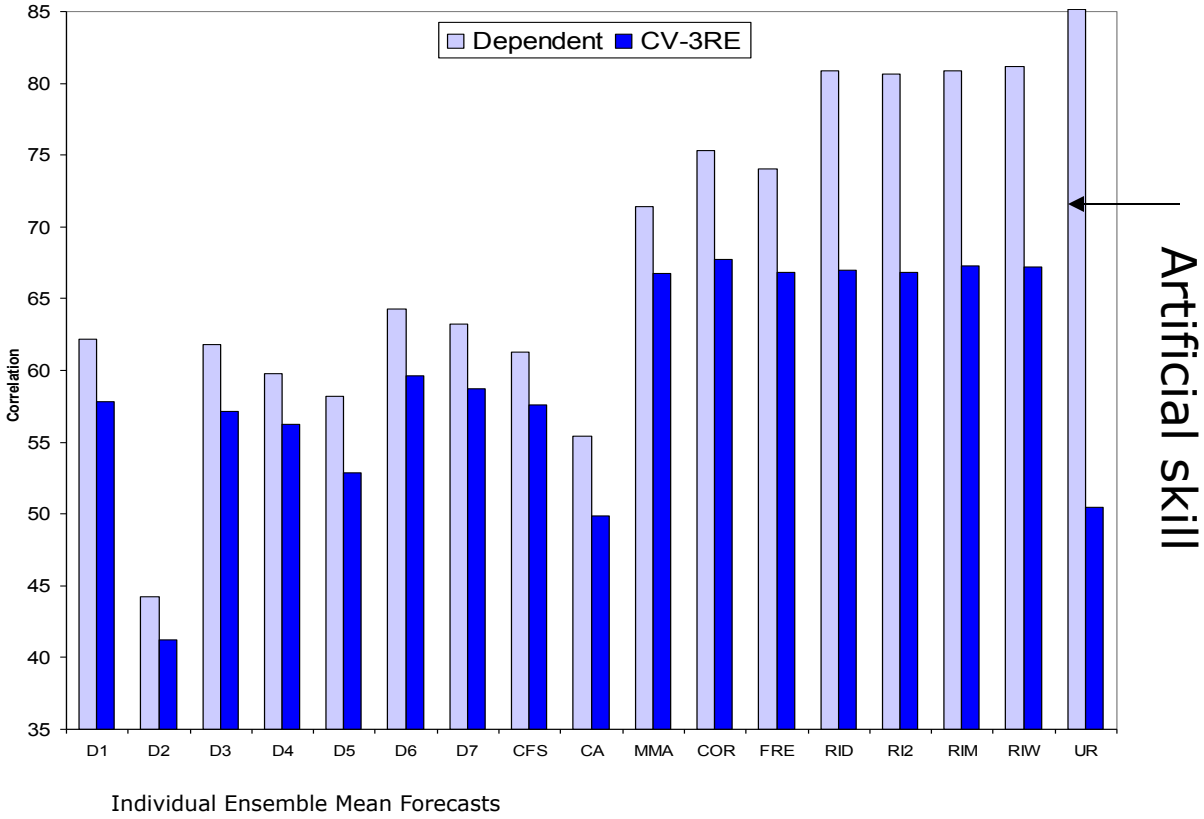
- Data: *Nine ensemble prediction systems (DEMETER+CFS+CA)*
  - At least 9 ensemble members per model
  - Hindcast length: Twenty-one years (1981-2001)
  - Monthly mean forecasts; Leads 0 to 5 months
  - Four initial month: Feb, May, Aug, Nov

Examples of Consolidation methods

Multi-model ensemble mean (MM)	$\alpha_i = 1/9$, $i=1,..K$, K number of methods
Correlation (COR)	$\alpha_i = \frac{\mathrm{cov}(\zeta_i, O)}{\sigma^2_{\zeta_i}}$, ζ_i time series forecast of *i-th* method
Frequency of best (FRE)	$\alpha_i = \{N_i/N\}$, N number of training years, $N_i = \left\{ \sum_N cases(\zeta_i) \mid \zeta_i = \min\{(\zeta_k - O)^2, k = 1,..K\} \right\}$
Ridging (RID)	$\vec{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}$ $\mathbf{A} = \mathbf{Z}^T\mathbf{Z}$, $\mathbf{b}=\mathbf{Z}^T\mathbf{O}$, λ is such that $\alpha_i \geq -0.01$, $i=1,..,K$ *and sum alpha squared small*
Double pass Ridging (RI2)	Set to zero any $\alpha_i < 0$, $i=1,...,K$ *after first RID pass.*
RID with MM constraint (RIM)	$\vec{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\left(\mathbf{b} + \frac{\lambda}{\mathrm{K}}\mathbf{1}\right)$
RID with weighted mean constraint (RIW)	$\vec{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}^*$ where $b_i^* = b_i\left(1 + \frac{\lambda}{a_{ii}f}\right)$ and $f = \sum_{i=1}^{K} \frac{b_i}{a_{ii}}$.
Unconstrained (UR)	$\vec{\alpha} = \mathbf{A}^{-1}\mathbf{b}$

Effect of cross-validation on multi-model combination methods

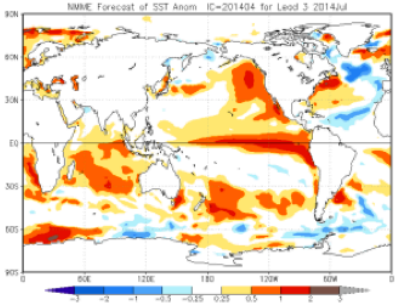
- D1, D2, ..D7 are distinct Ensemble predictions systems
- CFS is the NCEP Climate Forecast System
- CA is the statistical Constructed Analog
- Evaluations with dependent data have much larger skill but a large portion of it is artificial



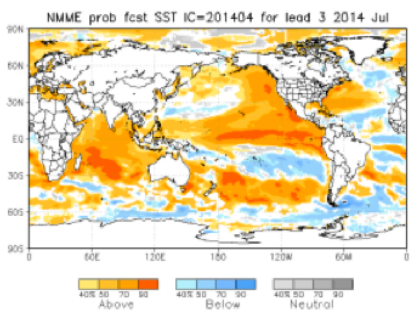
Current trend: Multi-model ensembles

Next steps:
a) Calibration
b) Combination methods
c) Probabilistic Verification

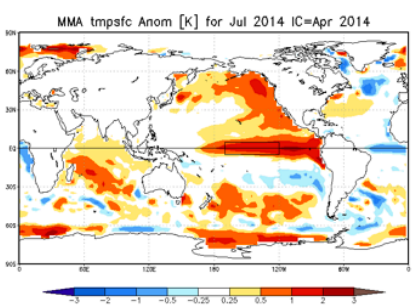
NMME



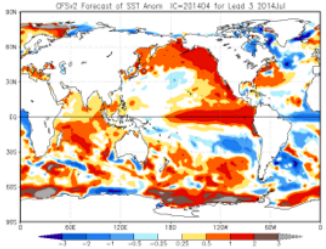
Prob fcst



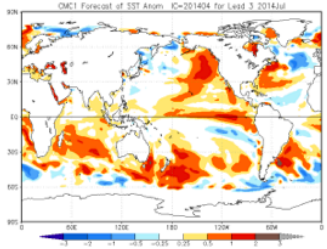
IMME



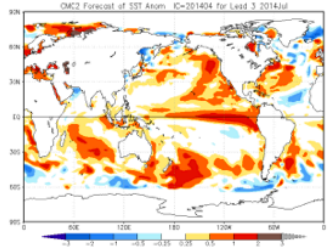
CFSv2



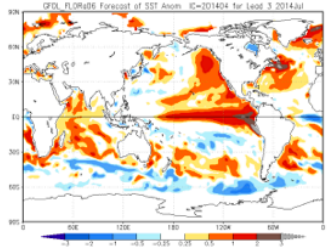
CMC1



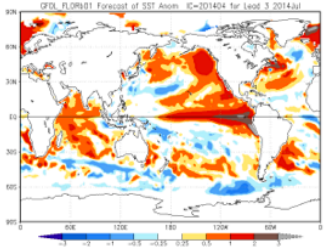
CMC2



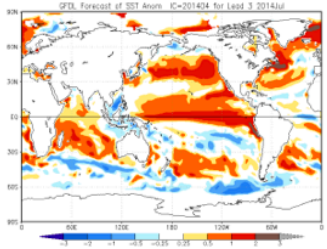
GFDL_FLORa06



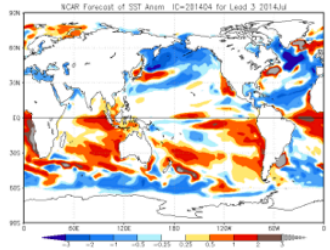
GFDL_FLORb01



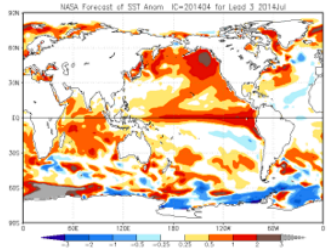
GFDL



NCAR



NASA



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