



Verification of Ensembles

Beth Ebert

Bureau of Meteorology Research Centre (BMRC),
Australia



Topics

- Verification philosophy
- Conventional verification methods and what they tell us
- Other approaches
 - Verifying individual events
 - Verifying "objects"
- Conveying forecast quality to users
- Sampling issues, including rare events
- Observations and observation errors
- TIGGE "standard" verification



Purposes of ensemble verification

User-oriented

- How accurate are the forecasts?
- Do they enable better decisions that could be made using alternate information (persistence, climatology, deterministic forecast)?

Intercomparison

- How do forecast systems differ in performance?

Calibration

- Assist in bias removal and downscaling

Diagnosis

- Pinpoint sources of error in ensemble forecast system

All purposes are relevant to TIGGE

← Operations ↔ Research →



What are we verifying?

Ensemble used to generate **probability distributions** with quality characterized by:

- **Skill** (accuracy) – are the forecasts close to the observed?
- **Spread** (variability) – does the forecast appropriately represent the uncertainty?

Which forecasts?

- All forecasts for a certain time period
 - Describes past performance
 - Usual operational verification
- All potential events
 - Estimate performance of forecast *system* for all possible weather, including rare events
 - Mainly done in research
- Individual event
 - Forecasters want results for most recent forecast
 - Users want to know forecast quality for certain significant events



Attributes of an ideal ensemble

Reliability

- Ability to give unbiased probability estimates for dichotomous (yes/no) forecasts
 - average frequency of occurrence equals forecast probability for all probability categories
- Forecast distribution represents distribution of observations
 - observations are statistically indistinguishable from ensemble members
 - on average, the spread of ensemble members equals the skill of the ensemble mean
- Reliability can be improved by calibration



Attributes of an ideal ensemble

Resolution

- Different probability forecasts correspond to different frequencies of observed events
 - forecast can be used to predict events and non-events
 - perfect resolution requires perfect deterministic forecasts (not possible for an ensemble, but should strive to maximize resolution)

Statistically speaking ... for dichotomous forecasts, the **reliability** and **resolution** fully describe the forecast quality.

Many samples required to describe reliability and resolution.



Conventional ensemble verification

"Old favourites" for probability forecasts:

- Reliability diagram
 - Histogram of forecast probabilities (sharpness diagram)
- Relative Operating Characteristic (ROC) diagram
 - ROC area
- Brier score, Brier skill score w.r.t. climatology

Also:

- Relative value

Methods for ensembles:

- Rank histogram (Talagrand diagram)
- Spread vs. skill

Deterministic verification:

- Verification of ensemble mean

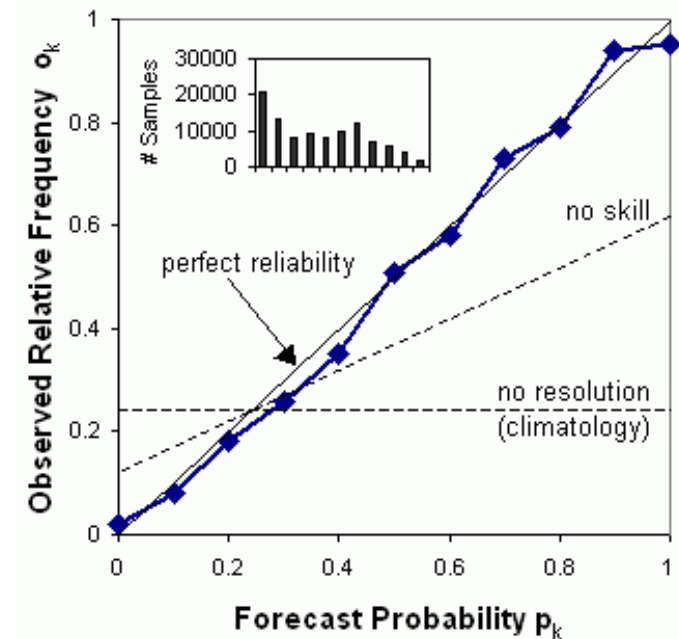
Reliability (attributes) diagram

Measures how well the predicted probabilities of an event correspond to their observed frequencies (reliability)

→ Plot observed frequency against forecast probability for all probability categories

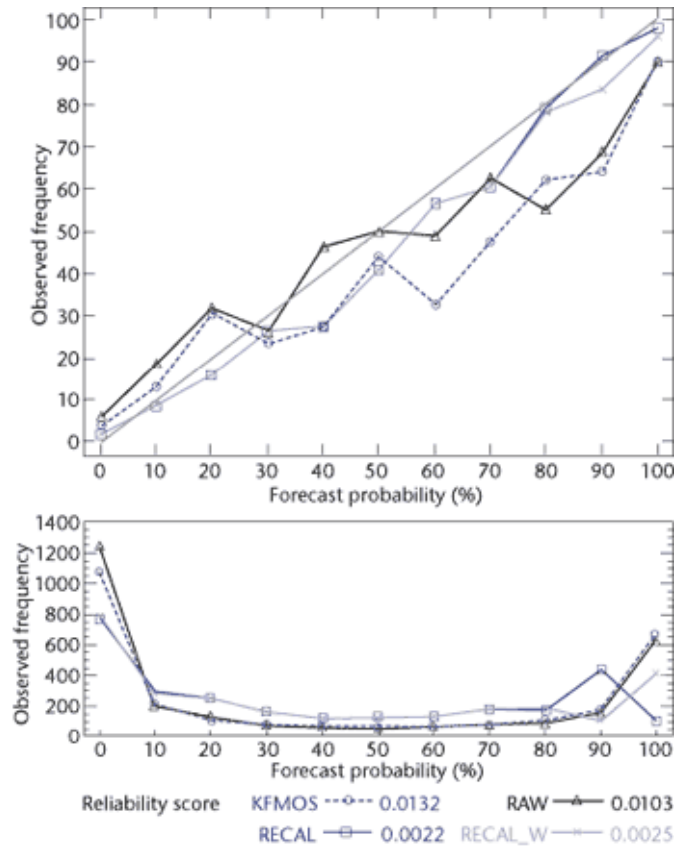
- Close to diagonal – good reliability
- Deviation from diagonal – conditional bias
 - Below diagonal – fcst probabilities too high
 - Above diagonal – fcst probabilities too low
 - Flatter curve – lower resolution

- Histogram of forecasts in each probability bin shows the sharpness of the forecast.
- The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast.



Reliability diagram

Example:



Reliability (top) and sharpness (bottom) diagrams for $T_{12} < 5^{\circ}\text{C}$ at $T+72$. Shades indicate the different levels of statistical processing applied as shown in the key.

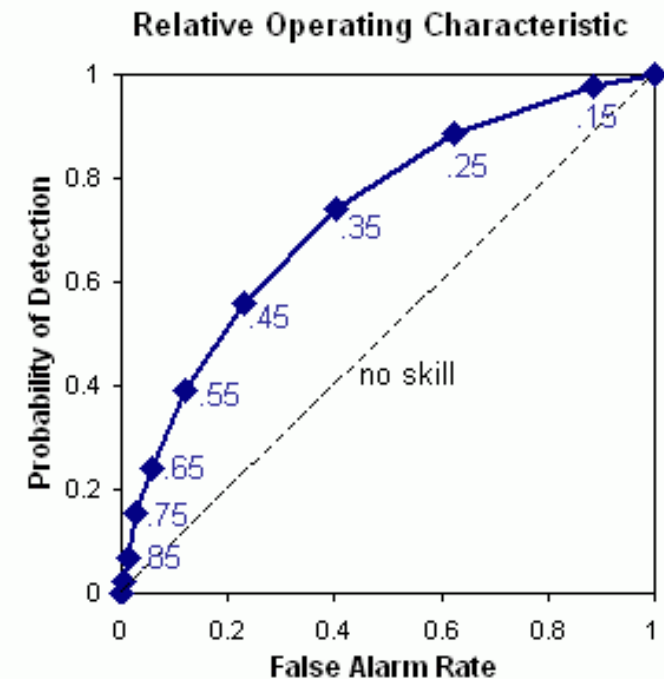
from "Verification of PREVIN site-specific probability forecasts", Met Office
(http://www.metoffice.com/research/nwp/publications/nwp_gazette/dec01/verif.html)

Relative Operating Characteristic (ROC)

Measures the ability of the forecast to discriminate between events and non-events (resolution)

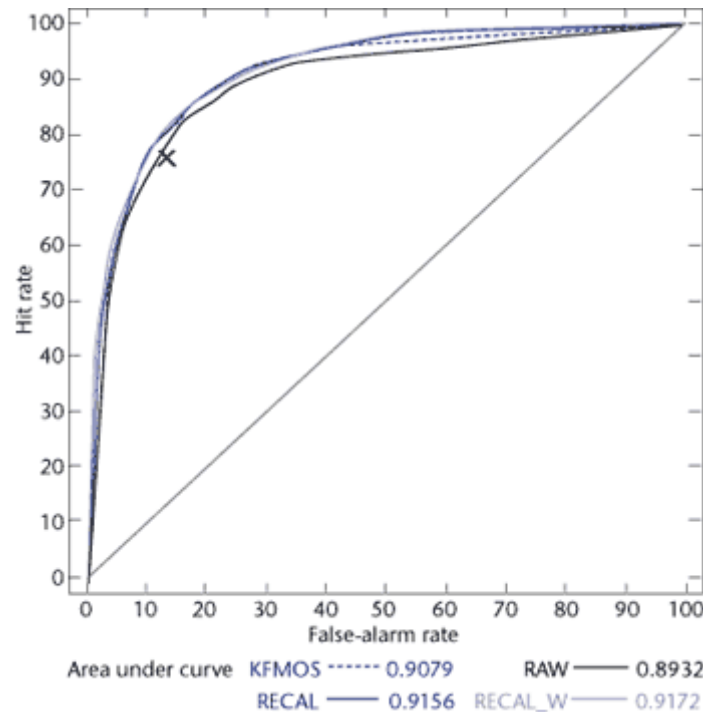
→ Plot hit rate H vs false alarm rate F using a set of varying probability thresholds to make the yes/no decision.

- Close to upper left corner – good resolution
 - Close to diagonal – little skill
- Area under curve ("ROC area") is a useful summary measure of forecast skill
 - Perfect: ROC area = 1
 - No skill: ROC area = 0.5
 - ROC skill score ROCS = $2(\text{ROC area} - 0.5)$
 - Not sensitive to bias.
 - The ROC is conditioned on the observations (i.e., given that Y occurred, what was the corresponding forecast?)
 - Reliability and ROC diagrams are good companions



Relative Operating Characteristic (ROC)

Example:



ROC diagram for $T_{12} < 5\text{ °C}$ at $T+72$. Shades indicate the different levels of statistical processing applied as shown in the key. The cross indicates the ROC (FAR, HR) of the ECMWF high-resolution deterministic model.

from "Verification of PREVIN site-specific probability forecasts", Met Office
(http://www.metoffice.com/research/nwp/publications/nwp_gazette/dec01/verif.html)



Brier (skill) score

Brier score measures the mean squared probability error

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

Murphy's (1973) decomposition into 3 terms (for K probability classes and N samples):

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

- Useful for exploring dependence of probability forecasts on ensemble characteristics
- Perfect score: 0 ← only possible for perfect deterministic forecast!

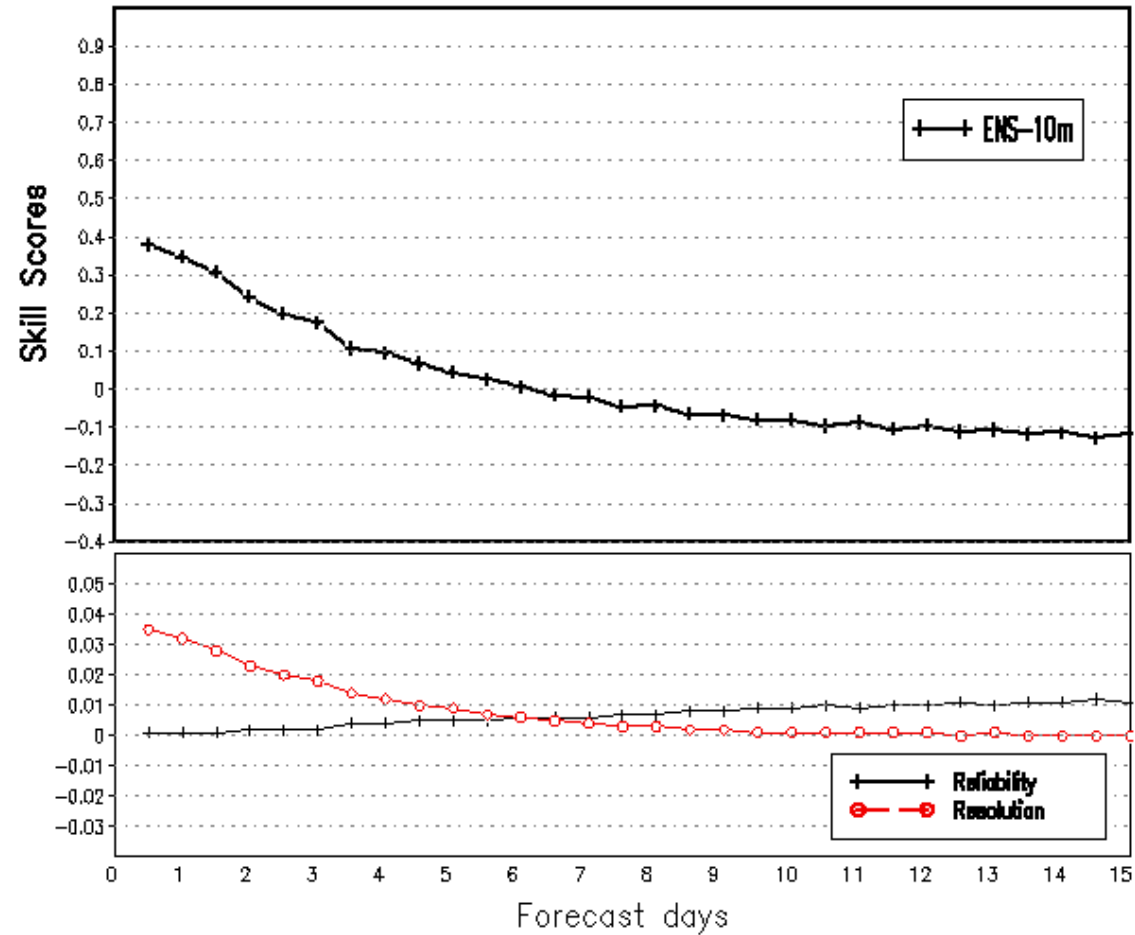
Brier *skill* score measures the relative skill of the forecast compared to climatology

$$BSS = 1 - \frac{BS}{BS_{clim}}$$

Brier skill score

Example

Northern Hemisphere 500 mb Height Brier Skill Scores (BSS)
Average For 20040701 – 20040731



from Y. Zhu, NCEP global ensemble verification
(http://wwwt.emc.ncep.noaa.gov/gmb/yzhu/html/opr/Z500_ROC_BSS.html)

Relative value score

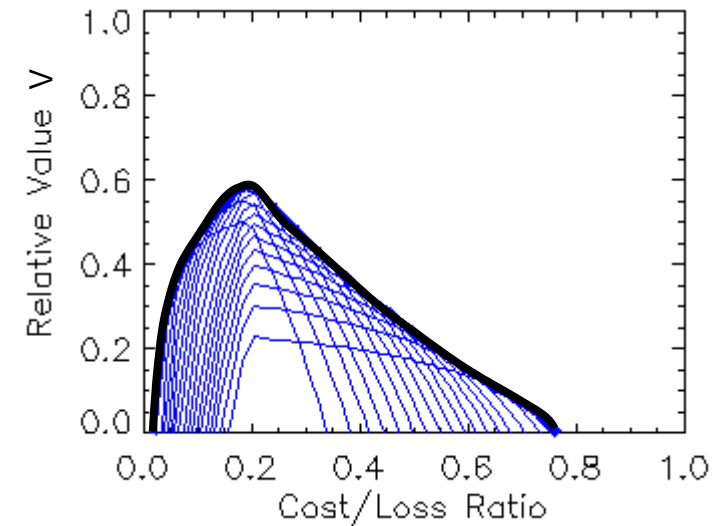
Measures the relative improvement in economic value as a function of the cost/loss ratio C/L for taking action based on a forecast as opposed to climatology

$$V = (1-F) - \left(\frac{1-C/L}{C/L} \right) \left(\frac{\bar{o}}{1-\bar{o}} \right) (1-H) \quad \text{if } C/L < \bar{o}$$

$$V = H - \left(\frac{C/L}{1-C/L} \right) \left(\frac{1-\bar{o}}{\bar{o}} \right) F \quad \text{if } C/L > \bar{o}$$

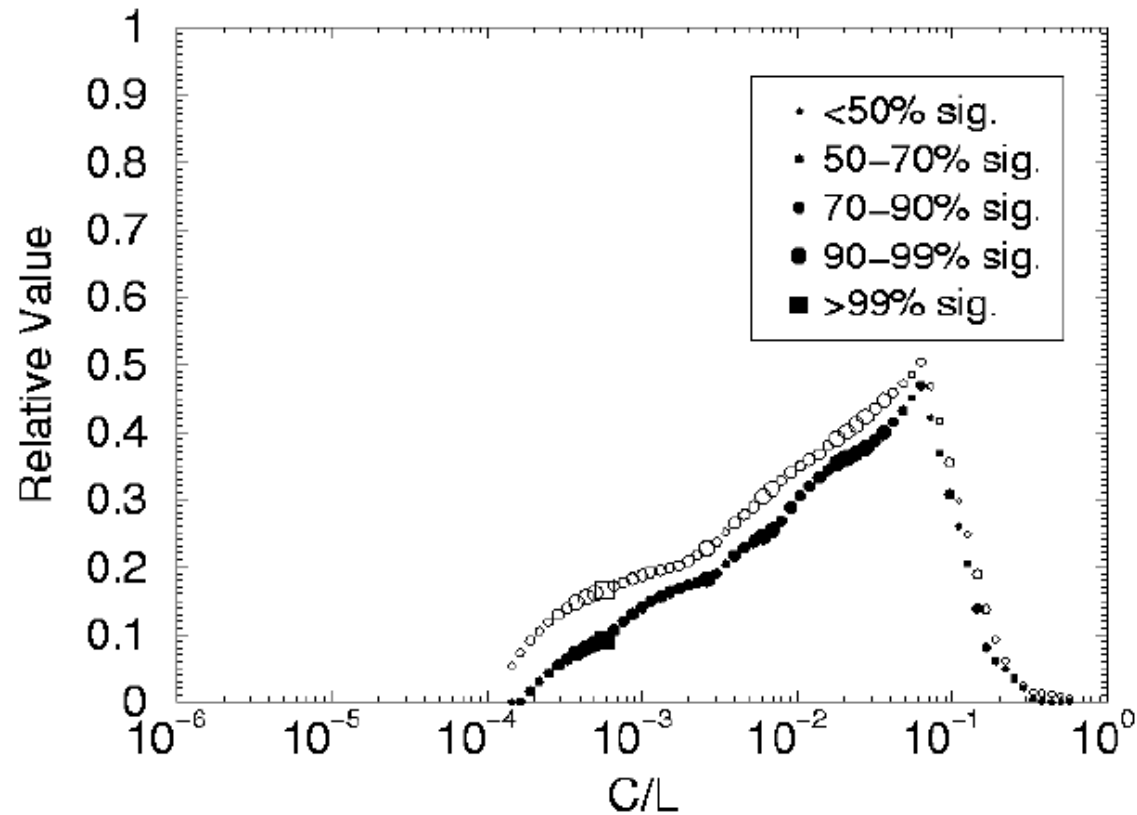
where H is the hit rate and F is the false alarm rate

- The relative value is a skill score of expected expense, with climatology as the reference forecast.
- Range: $-\infty$ to 1. Perfect score: 1
- Plot V vs C/L for various probability thresholds. The envelope describes the potential value for the ensemble system.



Relative value score

Example:

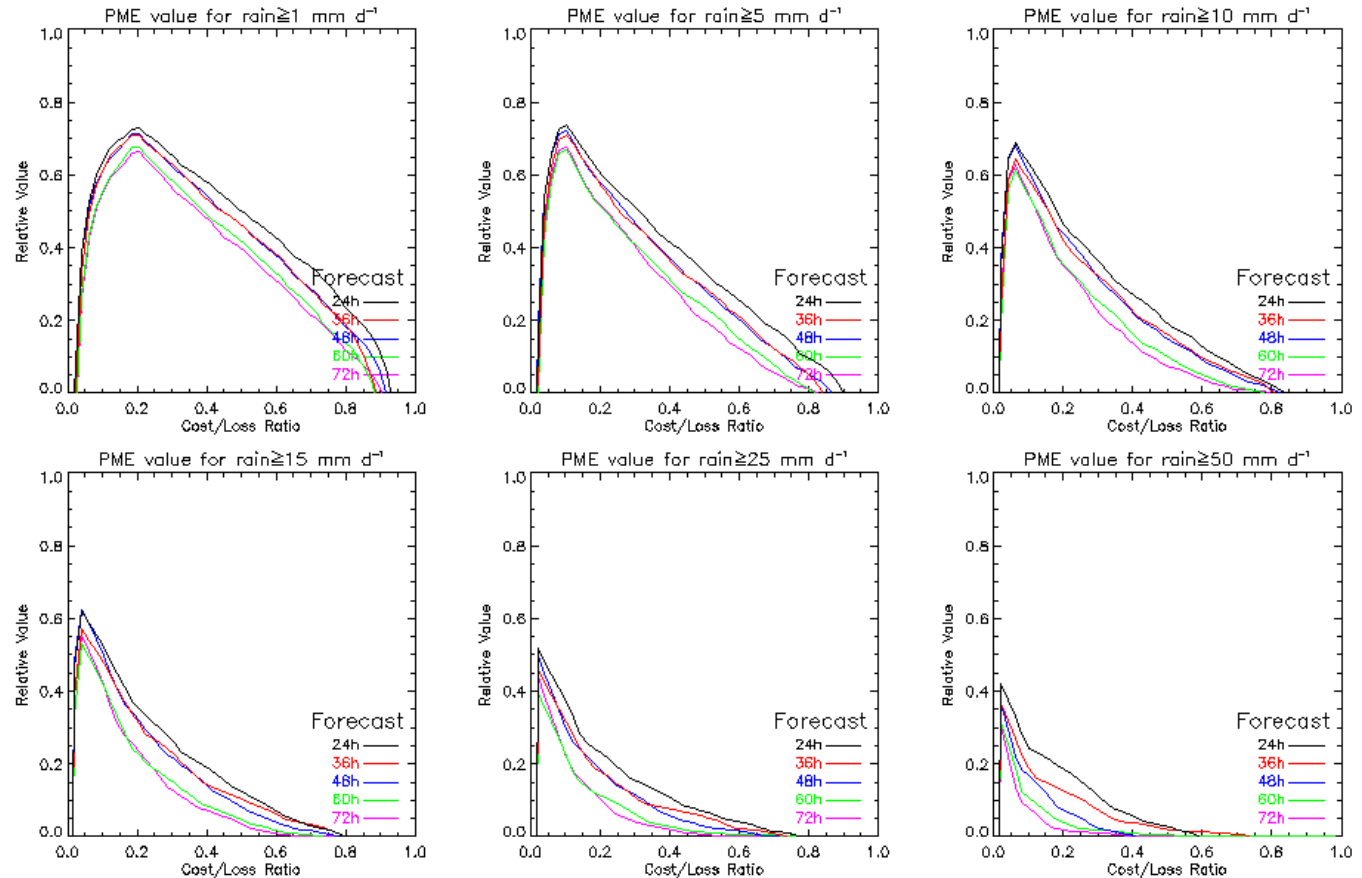


Relative value for 84h forecasts of 12-hour precipitation accumulation from the ECMWF ensemble (open circles) and T159 model (closed circles) based on spatial multi-event contingency tables.

From Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.*, **8**, 401-417.

Relative value score

Example:

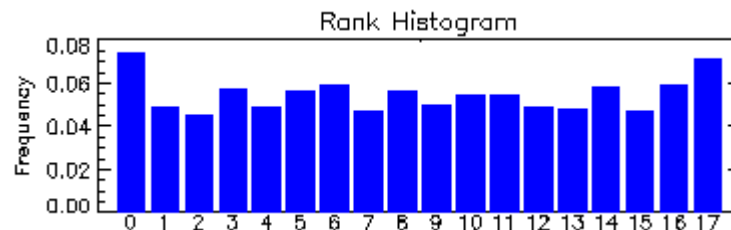


Relative value of Poor Man's Ensemble forecasts of precipitation over Australia during DJF 2004-05.

Rank histogram (Talagrand diagram)

Measures how well the ensemble spread of the forecast represents the true variability (uncertainty) of the observations

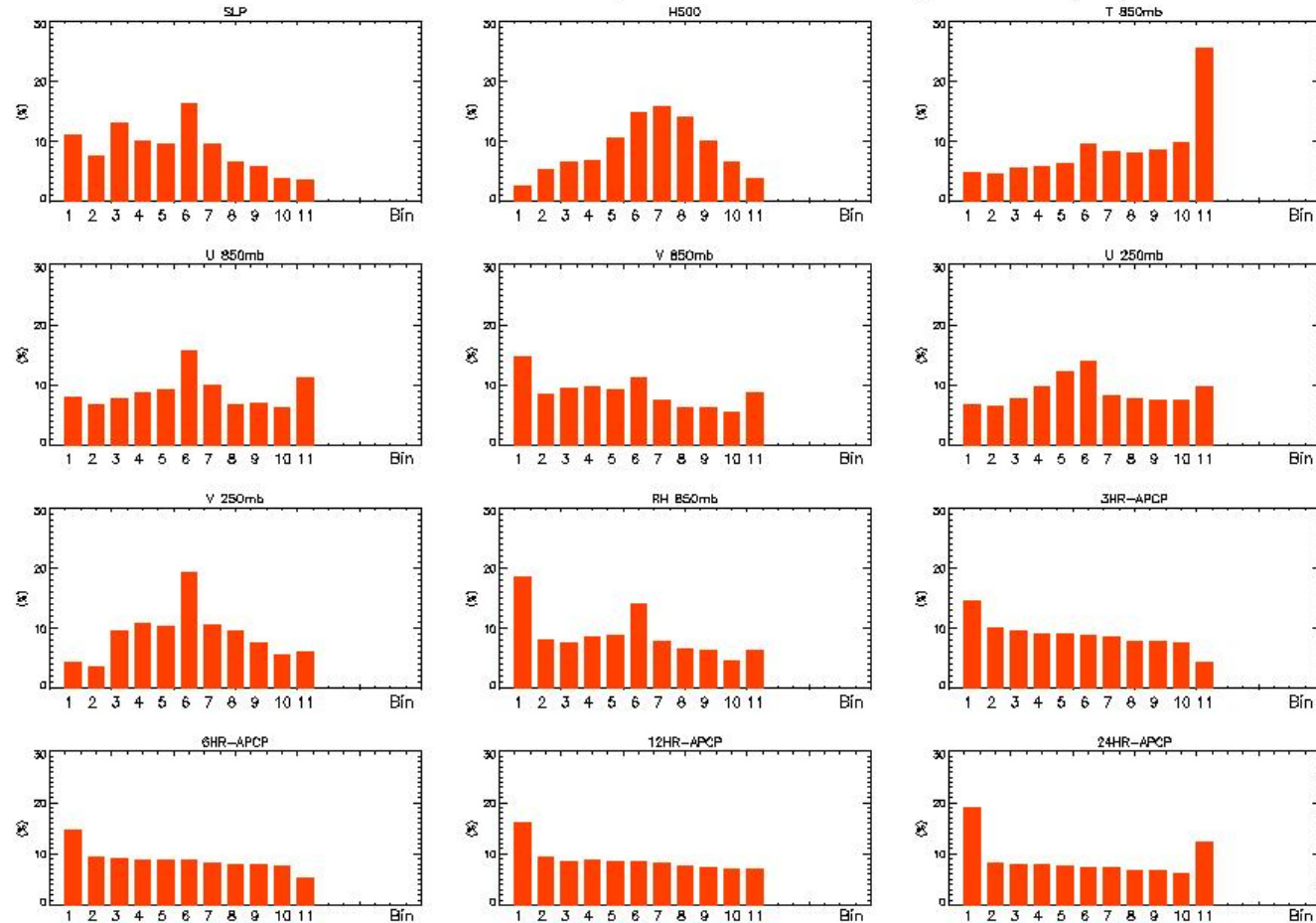
- Count where the verifying observation falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point.
- In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members.
 - Flat - ensemble spread correctly represents forecast uncertainty
 - U-shaped - ensemble spread too small, many observations falling outside the extremes of the ensemble
 - Dome-shaped - ensemble spread too large, too many observations falling near the center of the ensemble
 - Asymmetric - ensemble contains bias
 - A flat rank histogram does not necessarily indicate a skilled forecast, it only measures whether the observed probability distribution is well represented by the ensemble.



Rank histogram (Talagrand diagram)

Example:

Chance Ensemble Encompasses Anl at 27h for COM, from 9z 4/10/2003

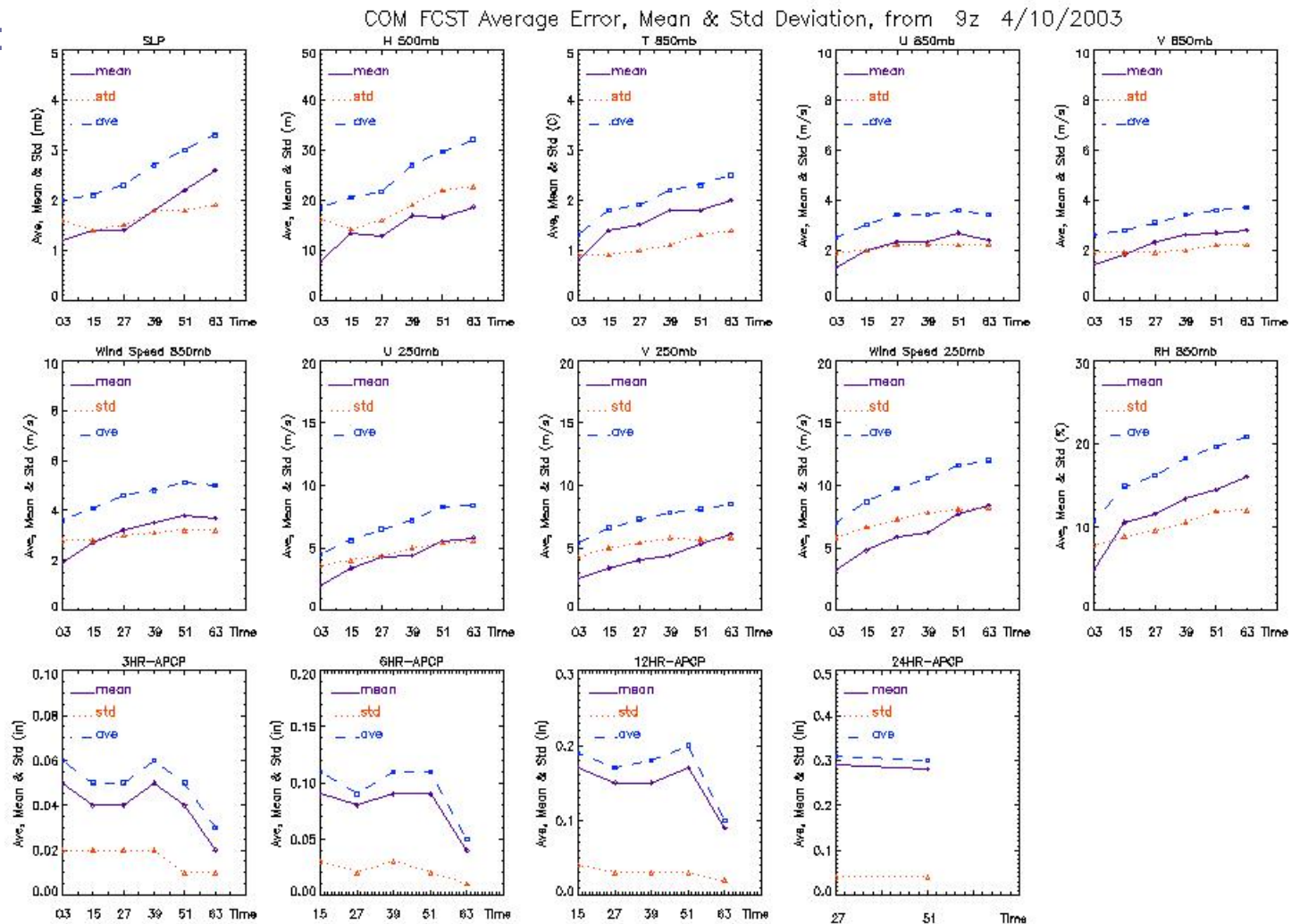


Note different spread and bias behaviour for different atmospheric variables!

SREF 27h forecasts (http://www.emc.ncep.noaa.gov/mmb/SREF/VERIFICATION/20030410_html/com_system_09z.html)

Spread – skill evaluation

Example:





Verification of ensemble mean

Debate as to whether or not this is a good idea:

Pros:

- Ensemble mean filters out smaller unpredictable scales, reflects model's skill
- Needed for spread – skill evaluation
- Forecasters and others use ensemble mean

Cons:

- Not a realization of the ensemble
- Different statistical properties to ensemble and observations

Scores:

- RMSE
- Anomaly correlation (AC)
- Other deterministic verification scores

Performance of ensemble mean should be compared to performance of control and hi-res forecasts



Who's using what?

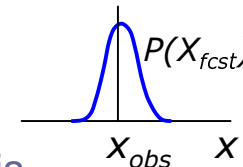
- WMO (ensemble NWP, site maintained by JMA)
 - Brier skill score, reliability diagram, economic value, ensemble mean & spread
- Some operational centers (ensemble NWP) – web survey

ECMWF	BSS, reliability diagram, ROC, ROC area, econ. value, spread/skill diagram
NCEP	RMSE and AC of ensemble mean, BSS, ROC area, rank histogram, RPSS, econ. value
Met Office	BSS, reliability diagram, ROC, rank histogram
BMRC	RMSE ensemble mean, BSS, reliability diagram, ROC, rank histogram, RPSS, econ. value

- DEMETER (multiple coupled-model seasonal ensemble) – see <http://www.ecmwf.int/research/demeter/d/charts/verification/>
 - Deterministic: anomaly correlation, mean square skill score, SD ratio
 - Probabilistic: reliability diagram, ROCS, RPSS
 - Economic value

Verifying individual events

- Forecasters and other users often want to know the quality of a forecast for a particular event
- Cannot meaningfully verify a single probability forecast
 - If it rains when the PoP was 30% was that a good forecast?
- ... but we can compare a probability distribution to a single observation
 - Want the forecast to be close to the observed (accurate), and sharp (not too much spread)
 - This approach implicitly assumes that the weather is *predictable* and the uncertainty comes from the forecast system
 - best used at short time ranges and/or large spatial scales
- Methods for individual or collections of forecasts
 - (Continuous) Ranked Probability Score
 - Wilson (1999) score
 - Ignorance

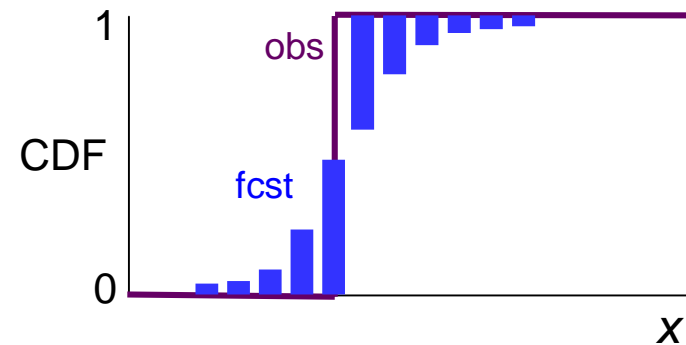


Ranked probability score

Measures the squared difference in probability space when there are multiple probability categories

$$RPS = \frac{1}{M-1} \sum_{m=1}^M (CDF_{fcst,m} - CDF_{obs,m})^2$$

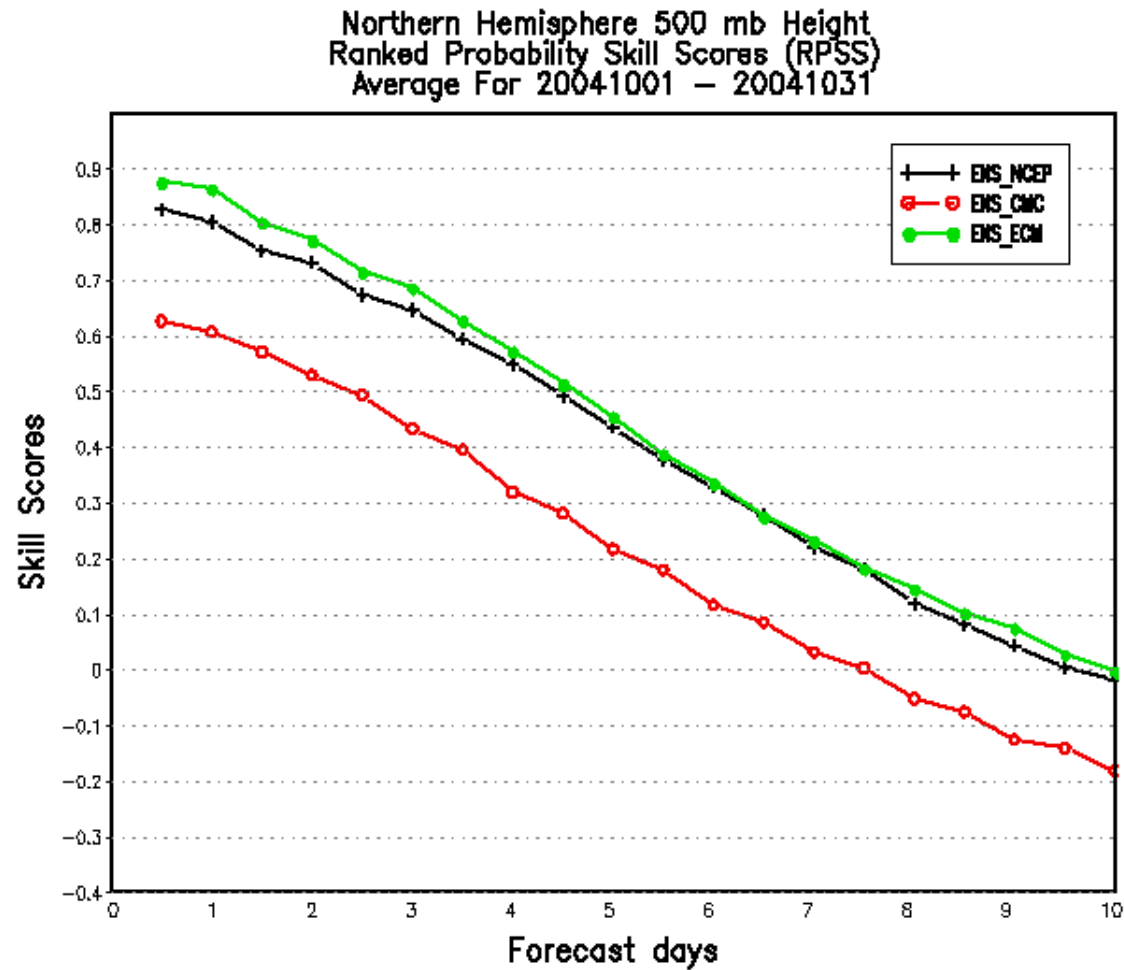
for M probability classes.



- Takes into account the ordered nature of the predicted variable (for example, temperature going from low to high values)
- Emphasizes accuracy by penalizing "near misses" less than larger errors
- Rewards small spread if the forecast is accurate
- Perfect score: 0
- RPS skill score w.r.t. climatology: $RPSS = 1 - \frac{RPS}{RPS_{clim}}$

Ranked probability skill score

Example:



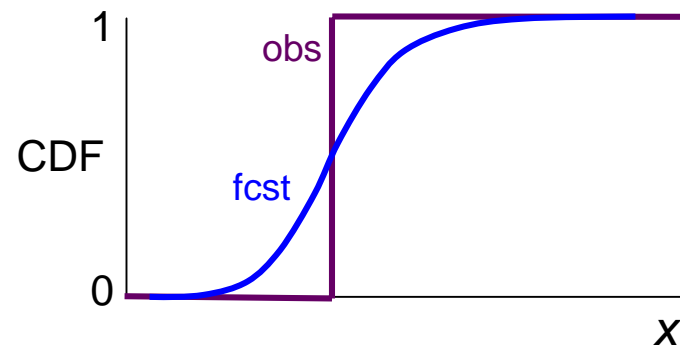
from Y. Zhu, NCEP global ensemble verification
(http://wwwt.emc.ncep.noaa.gov/gmb/yzhu/html/opr/Z500_ROC_BSS.html)

Continuous ranked probability score

Continuous ranked probability score (CRPS) measures the difference between the forecast and observed CDFs

$$CRPS = \int_{-\infty}^{\infty} (P_{fcst}(x) - P_{obs}(x))^2 dx$$

- Same as Brier score integrated over all possible threshold values
- Same as Mean Absolute Error for deterministic forecasts
- Advantages:
 - sensitive to whole range of values of the parameter of interest
 - does not depend on predefined classes
 - easy to interpret
 - has dimensions of the observed variable
- Rewards small spread (sharpness) if the forecast is accurate
- Perfect score: 0



Wilson (1999) score

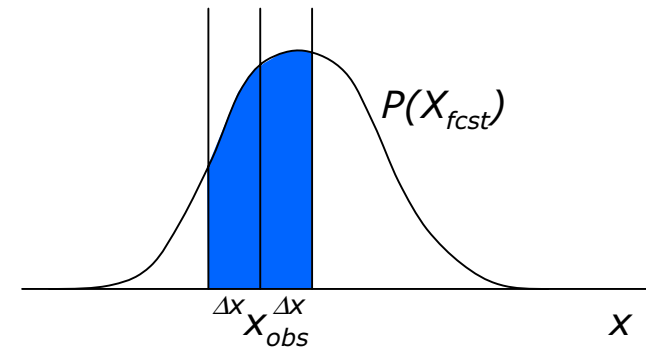
Measures the accuracy of the forecast probability distribution

$$WS = P(x_{obs} | X_{fcst}) = \int_{X_{obs} - \Delta X}^{X_{obs} + \Delta X} P(X_{fcst}) dx$$

for some "acceptable" range Δx

- Advantages:
 - user-oriented
 - simple, understood as a probability
- Like CRPS, rewards accuracy and sharpness
- Perfect score: 1
- Accounting for climatological variability:
 - choose Δx as a fraction of climatological variance
 - skill score with respect to climatology

$$WSS = \frac{WS - WS_{clim}}{1 - WS_{clim}}$$

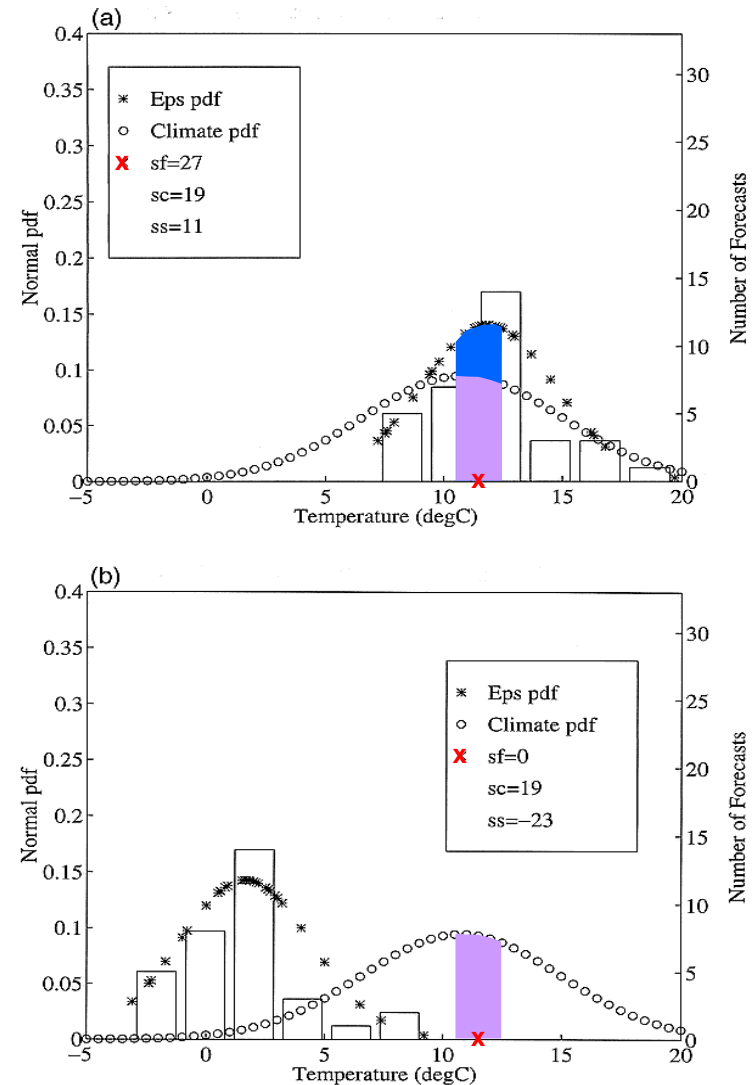


Wilson (1999) score

Example:

Histogram of ECMWF ensemble temperature forecasts, fitted normal distribution (stars), and climatological distribution (circles) for (a) 72-h and (b) 168-h projections valid 17 May 1996 for Toronto, ON, Canada (PIA). The observed temperature is indicated by the cross on the abscissa, and the window for a correct forecast is $\Delta T = \pm 1\text{C}$. The probability score for the forecast is “sf,” “sc” is the probability score for climatology, and “ss” is the skill score for this case. Score values are multiplied by 100.

(from Wilson et al., 1999)





Ignorance score

Measures the amount of data compression required the forecast to represent the truth

$$IGN = -\log_2 p_{k_{obs}}$$

for a categorical probabilistic forecast defined by p_k ($k=1, \dots, K$)

- Advantages:
 - strictly proper – discourages hedging
 - makes no assumption about the shape of the PDF
 - can be used with rank histograms also
- Rewards accuracy and sharpness
- Perfect score: 0

Can specify an ignorance skill score w.r.t. climatology
if K equi-probable categories are used:

$$ISS = 1 - \frac{\log_2 p_{k_{obs}}}{\log_2 K}$$

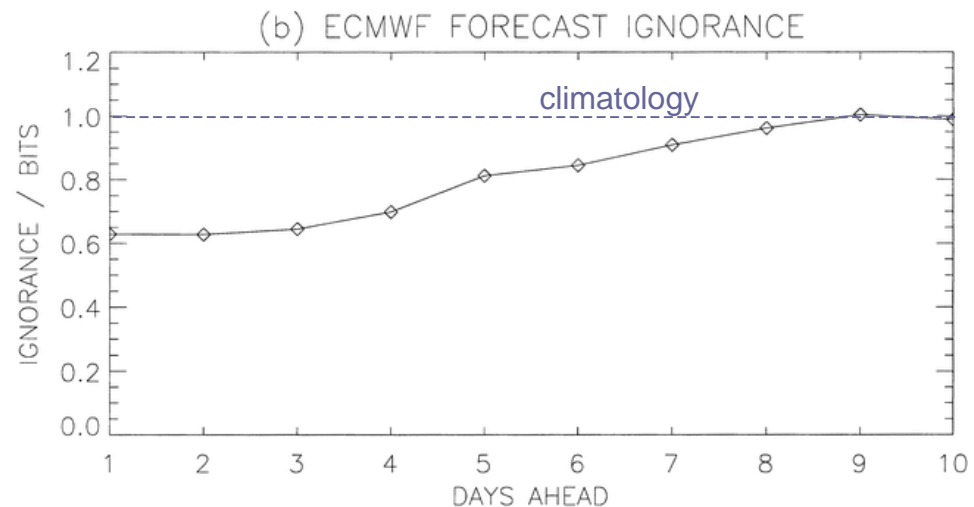
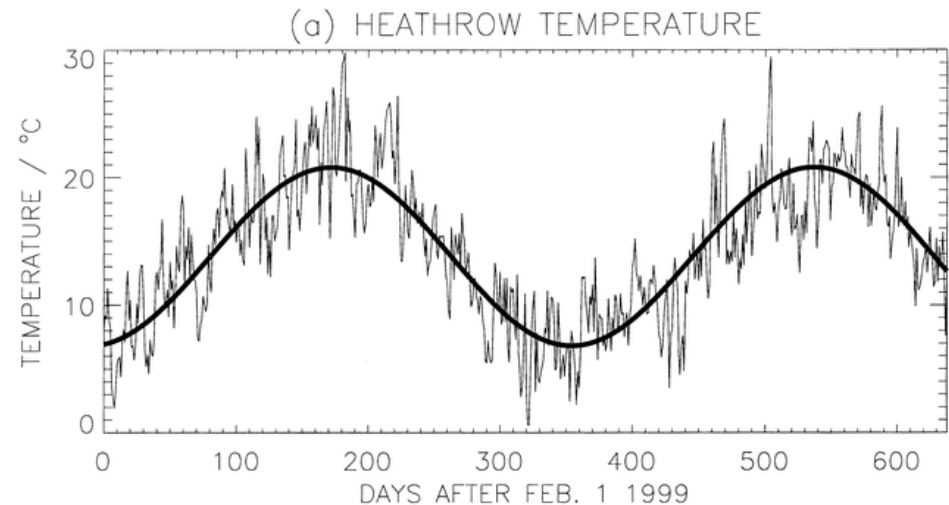
Ignorance score

Example:

The observed temperature at London's Heathrow airport (thin line) and an average seasonal cycle (thick line).

The average ignorance of probabilistic forecasts of whether the temperature will be above or below the seasonal average. The daily forecasts were constructed using operational 51-member ECMWF ensembles.

(from Roulston and Smith, 2002)



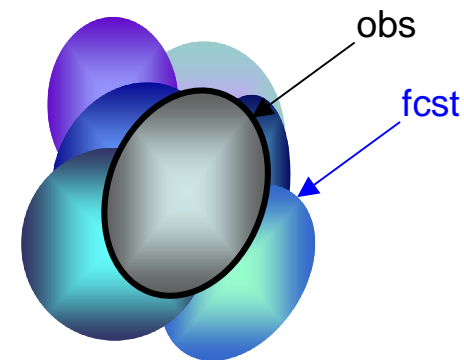
Verifying "objects"

Significant weather events can often be viewed as 2D objects

- tropical cyclones, heavy rain events, deep low pressure centres
- objects are defined by an intensity threshold

What might the ensemble forecast look like?

- spatial probability contour maps
- distributions of object properties
 - location, size, intensity, etc.



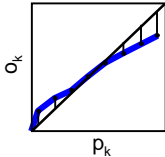
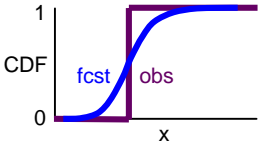
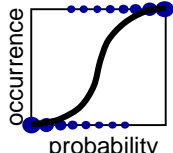
Strategies for verifying ensemble predictions of objects

- Verify spatial probability maps
- Verify distributions of object properties
 - many samples – use probabilistic measures
 - individual cases – CRPS, WS, IGN
- Verify ensemble mean
 - spatially average forecast objects
 - generated from average object properties

Conveying forecast quality to users

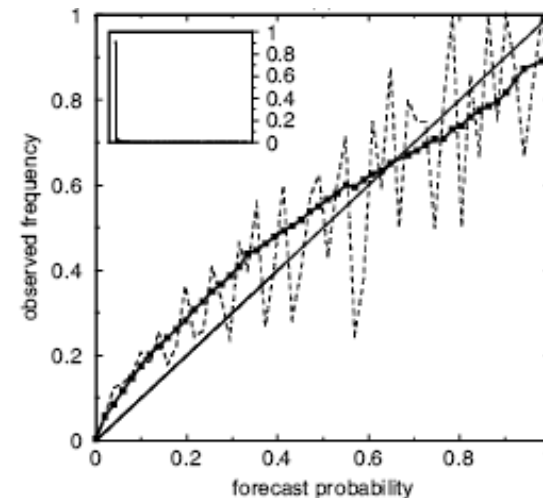
Forecasters and other users are ~comfortable with standard verification measures for deterministic forecasts

Are there similar easy-to-understand measures for probabilistic forecasts?

Deterministic	Probabilistic (suggestions)	Visual aid
Mean bias	Reliability term of BS $\frac{1}{N} \sum_{k=1}^K n_k (p_k - \bar{o}_k)^2$	
RMS error	Brier score (square root) $\sqrt{BS} = \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2}$	
Mean absolute error	CRPS $\int (P_{fcst}(x) - P_{obs}(x))^2 dx$	
Correlation	R^2 for logistic regression	

Sampling issues – rare events

- Rare events are often the most interesting ones!
- Coarse spatial resolution may not capture intensity of experienced weather
- Forecast calibration approaches – see Tom Hamill's talk
- Difficult to verify probabilities on the "tail" of the PDF
 - Too few samples to get robust statistics, especially for reliability
 - Finite number of ensemble members may not resolve tail of forecast PDF
- An approach for improving robustness of verification:
 - Fit ROC for all events (incl. rare) using bi-normal model, then relate back to reliability (Atger, QJRMS, 2004) to get *estimated* forecast quality for under-sampled categories
 - Fitted reliability also be used instead of "raw" frequencies to calibrate ensemble



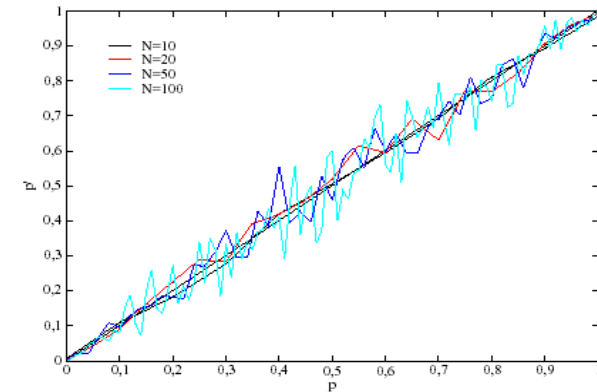
Sampling issues – size of ensemble vs number of verification samples

Robustness of ensemble verification results depends both on the size of the ensemble and the number of verification samples

For an ensemble with N members, and a verification sample of M realizations:

- Increasing N without increasing M improves the resolution but degrades the reliability
- If we wish to know the reliability to a precision ε , need sample size of

$$M \geq \frac{2}{\varepsilon^2} N \ln N$$



For $\varepsilon = 10\%$

N	5	10	20	50	100	1000
$M \geq$	1963	5549	14087	44690	103447	1.5×10^6

Candille and Talagrand, 2004: On limitations to the objective evaluation of ensemble prediction systems. *Workshop on Ensemble Methods, Exeter, October 2004.*

Stratification of samples

- Verification results vary with region and season
- Inhomogeneity in sample populations leads to overestimates of forecast skill
 - Example: Verification of ensemble forecasts for tropical rain
 - using 1 year of data to get lots of samples → great results!
 - at least some of the "skill" simply reflects wet season vs dry season
- Stratify data into homogeneous sub-samples
 - Must have enough samples to give robust statistics
 - If we wait too long then the model is changed!



Uncertainty of verification results

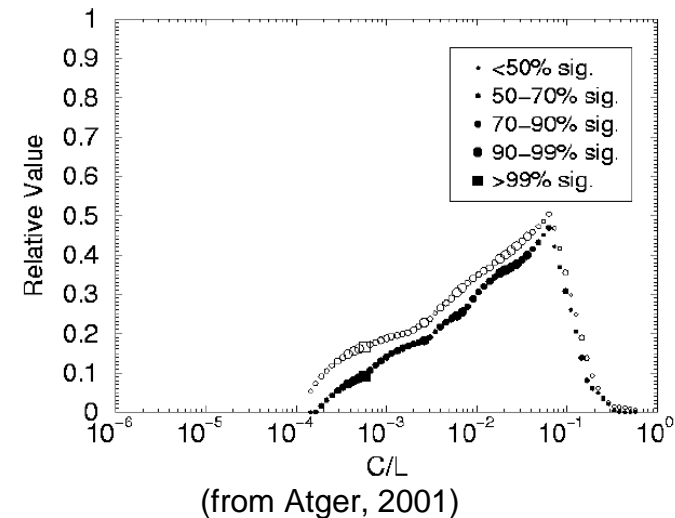
Are the ensemble forecasts significantly better than random chance?

Does ensemble A perform significantly better than ensemble B?

Significance levels and/or confidence intervals address these questions

Non-parametric resampling (Monte Carlo, bootstrap) methods easy to use

- Hamill (1999) approach
 - build null distribution by repeated (1000+) random sampling from collected data
 - assess significance of test result by where it falls in null distribution
- Basic bootstrap
 - score 1000+ sample sets generated using random draw (with replacement)
 - determine confidence intervals from distribution of sample scores





Effects of observation errors

Observation errors add uncertainty to the verification results

- True forecast skill is unknown
 - An imperfect model / ensemble may score better!
- Extra dispersion of observation PDF

Effects on verification results

- RMSE – overestimated
- Spread – more obs outliers make ensemble look under-dispersed
 - Saetra et al (2004) compensate by adding obs error to ensemble
- Reliability – poorer
- Resolution – greater in BS decomposition, but ROC area poorer
- CRPS, WS, IGN – poorer mean values

Can we remove the effects of observation error?

- More samples helps with reliability estimates
- Error modeling – study effects of applied observation errors
- Need "gold standard" to measure actual observation errors

Not easy!



Sources of observation data

- Surface weather (temperature, precipitation, etc.)
 - Measurements at sites
 - "pure" observations, experienced by public
 - errors of representativeness, scale mismatch
 - most appropriate for verifying downscaled forecasts
 - Gridded analyses
 - more representative of model scale
 - even spatial distribution of observations
 - analysis process introduces errors
- Upper level fields (Z_{500} , T_{850} , etc)
 - Gridded analyses
 - verification against model's own analysis is incestuous
 - TIGGE – consider using multi-model analysis for ensemble intercomparison and verification



TIGGE verification "standards"

1999 workshop on "Ensemble Forecasting in the Short to Medium Range" (Hamill et al., 2000) recommended a standard suite of verification scores and diagrams:

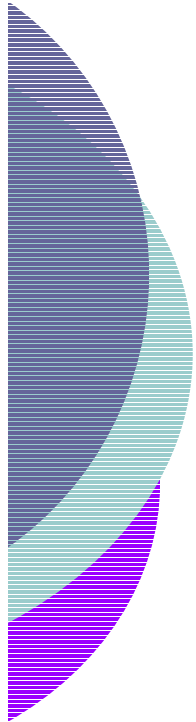
- Probabilistic scores: BSS, RPS/RPSS
 - Diagrams: Reliability, ROC, rank histograms
- with more emphasis on sensible weather

WMO standards for comparing performance of ensemble NWP:

- Deterministic scores: Ensemble mean, ensemble spread
- Probabilistic scores: BSS
- Diagrams: Reliability, economic value
- Atmospheric variables: PMSL, Z_{500} , $|V|_{850}$, T_{850} , 24h precipitation

What do we want for TIGGE?

- All of the above, or some optimal subset?
- Encourage experimentation with user-oriented, object-oriented, and other new verification methods



Thank you!