Probability and Ensemble Methods

Laurence J. Wilson
DTC Verification Workshop
Boulder, 16/4/08
Outline

• Introduction – Principles of verification
• Murphy’s attributes
• Probability interpretation from ensembles
• Survey of probability and eps verification methods
• Spatial verification of ensembles?
• Summary
Principles of (Objective) Verification

• Verification activity has value only if the information generated leads to a decision about the forecast or system being verified
  – User of the information must be identified
  – Purpose of the verification must be known in advance
  – Better to ask the user what he wants to know about the product than “what verification he wants to include”
  – If for more than one user group, might need flexible or multiple designs.

• No single verification measure provides complete information about the quality of a forecast product.

• Forecast must be stated in such a way that it can be verified
<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DEFINITION</th>
<th>RELATED MEASURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bias</td>
<td>Correspondence between mean forecast and mean observation</td>
<td>bias (mean forecast probability-sample observed frequency)</td>
</tr>
<tr>
<td>2. Association</td>
<td>Strength of linear relationship between pairs of forecasts and observations</td>
<td>covariance, correlation</td>
</tr>
<tr>
<td>3. Accuracy</td>
<td>Average correspondence between individual pairs of observations and forecasts</td>
<td>mean absolute error (MAE), mean squared error (MSE), root mean squared error, Brier score (BS)</td>
</tr>
<tr>
<td>4. Skill</td>
<td>Accuracy of forecasts relative to accuracy of forecasts produced by a standard method</td>
<td>Brier skill score, others in the usual format</td>
</tr>
</tbody>
</table>
Comments on summary scores

• Brier score (2 category) and RPS (multicategory) are basic accuracy scores for probability forecasts

• Skill – recommend long term climatology for stations be included. Skill vs sample climatology is usually done, but not a “fair” measure for some users.
<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DEFINITION</th>
<th>RELATED MEASURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Reliability</td>
<td>Correspondence of conditional mean observation and conditioning forecast, averaged over all forecasts</td>
<td>Reliability component of BS, MAE, MSE of binned data from reliability table.</td>
</tr>
<tr>
<td>6. Resolution</td>
<td>Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts.</td>
<td>Resolution component of BS</td>
</tr>
<tr>
<td>7. Sharpness</td>
<td>Variability of forecasts as described by distribution of forecasts</td>
<td>Variance of forecasts</td>
</tr>
<tr>
<td>8. Discrimination</td>
<td>Difference between conditional mean forecast and unconditional mean forecast, averaged over all observations</td>
<td>Area under ROC, measures of separation of conditional distributions; MAE,MSE of scatter plot, binned by observation value</td>
</tr>
<tr>
<td>9. Uncertainty</td>
<td>Variability of observations as described by the distribution of observations</td>
<td>Variance of observations</td>
</tr>
</tbody>
</table>
Reliability, Resolution and Sharpness

Reliability:
The correspondence between the conditional observed frequency for each forecast probability

Resolution:
-The variance of the conditional observed frequencies about the climatological frequency, conditioned on the forecast
-A component of the Brier score
-Steeper slope than 45 degree line suggests over-resolved forecasts.

Sharpness:
The variance of the forecasts
Comments on Reliability Table

- Essential part of verification of probability forecasts
- Needs a large verification sample – controls on binning of data by forecast
- Conditioning by forecast means that gives information forecasters (or modelers can use) – calibration, or other alteration of forecast strategy
- Relates to Brier score
**Discrimination**: The tendency for the distribution of forecast probabilities when the event occurs to differ from the distribution when the event doesn’t occur.

**Uncertainty**: The variance of the observations. If binary then the max is 0.25

A component of the Brier score.

Measured by the ROC and its distance statistics.

![Illustration of discrimination](image)
**ROC example - 24 h POP (>1 mm)**

The Likelihood diagram shows the two conditional probability distributions.

The distance can be computed directly and is given in terms of the std of the distribution for non-occurrences.
Comments on the ROC

- Computation methodology different depending on circumstances – will need two different methods (at least) and lots of guidance on appropriate use
- Essential – related to value studies
- Since conditional on obs, evaluates the forecast from decision maker’s perspective
Estimating probabilities from ensembles

• Probability, pdf and cdf estimates are interpretations of eps output
• Three kinds:
  – Discrete (empirical)
  – Histogram
  – Continuous (parametric or non-parametric)
PDF interpretation from ensembles

Discrete

"Histogram"

pdf

1
cdf

P(x)

< X

0

P(x)

< X

x →

x →
Why density fitting is needed

• For extreme event prediction, to estimate centile thresholds.
• For some scoring rules e.g. ignorance
• Assists with the ROC computation.
• Kernel density estimation
  – Is non-parametric
  – Amount of smoothing determined by the bandwidth
  – Gaussian kernels fine for unbounded variables; gamma kernels for precip.
Examples of distribution fitting (Peel and Wilson, 2008)
Kernel Density fitting of ensemble cdfs
Survey of verification methods for ensembles

- Evaluate the ensemble distribution
  - Rank Histogram
  - CRPS, CRPSS (Hersbach, 2000)
- Evaluate the ensemble distribution in the vicinity of the observation
  - Wilson et al, 1999
  - Ignorance score (Roulston and Smith, 2002)
- Evaluate probabilities from the ensemble distribution
  - Brier score (accuracy), reliability, resolution
  - Reliability (attributes) diagrams
  - ROC area (discrimination)*
  - BSS, RPSS (skill)
Quantification of “departure from flat”

\[
\text{RMSE} = \sqrt{\frac{1}{N+1} \sum_{k=1}^{N+1} \left( S_k - \frac{M}{N+1} \right)^2}
\]

\[
\sqrt{\frac{MN}{(N+1)^2}}
\]
Comments on Rank Histogram

• Can quantify the “departure from flat”
• Not a “real” verification measure
• Who are the users?
Continuous Rank Probability Score

CRPS - 40 day training period
Comparison with Gaussian - b1

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

-difference between observation and forecast, expressed as cdfs
-defaults to MAE for deterministic fcst
-flexible, can accommodate uncertain obs
Probability score

- The probability assigned by the ensemble in the vicinity of the observation
- Maximized for sharp forecasts, correctly positioned
- Can be used to evaluate one forecast
- Not strictly proper

![Diagram showing probability density over temperature (°C)]
Ignorance Score (Roulston and Smith, 2002)

- From information theory, the number of bits needed to transmit the probability of the verifying category.
- \( \text{IGN} = -\log_2(f_i) \) where \( f_i \) is the probability assigned to the verifying category.
- Goes to infinity for 0 probability.
- Heavily penalizes low probabilities.
- Similar to probability score in that it considers the verification in the vicinity of the observation only.
Properness

• (Strictly) Proper: Forecaster achieves best score (only) if he issues a forecast which corresponds with his true belief

• Importance debatable in practice
  – Assumes one would “cheat” if one could
  – BUT – needed so that comparison between competing forecasts identifies the more accurate.
Data and method

- **Data**
  - 3.5 years of ensemble forecasts of precipitation from 36 Canadian stations, 24h accumulations, 0 to 10 days
  - Corresponding observations quality controlled without reference to models
  - Verification sample stratified into warm and cool seasons
  - Long-term precipitation climatology (~30 years) for all 36 stations as distribution

- **Method**
  - Using the long-term climatology, find 90th, 95th and 99th percentile thresholds for each station.
  - E.g. 90th percentile for Vancouver is 14.4 mm
  - Probability of exceedence of these thresholds as estimated from the ensemble forecast distribution (gamma distributions)
Brier Skill scores for eps forecasts (Peel and Wilson, 2008)
Probability of week long dry period
Spatial verification methods for ensembles – some thoughts

• Verification of centers (hurricanes, low centers, precip maxima etc)
  – Ideal application of the minimum spanning tree (Smith 2001, Wilks 2004)
  – Evaluates whether the ensemble “contains” the observation spatially in 2-d

• Other ideas? Bill Gallus?
  – How to match spatial (deterministic obs) with field of probability forecasts?
Minimum Spanning Tree – MST (Wilks 04)
Summary

• Importance of taking user requirements into account in the design of a verification system

• System for ensembles should include the capability to fit distributions with a non-parametric method (especially for extreme “high impact weather” verification, also for ROC)

• Spatial verification using MST
Thank you!
Ensemble verification involves comparing single observations with ensemble distributions, or at least, multiple forecasts.

What is a perfect ensemble forecast?

Is reliability enough?

Reliability: “For all instances where a pdf \( f(x) \) is forecast, the distribution of observations is equal to \( f \)”