Verification of ensembles

Barbara G. Brown

Acknowledgments: Tom Hamill, Laurence Wilson, Tressa Fowler
How good is this ensemble forecast?
Questions to ask before beginning?

• How were the ensembles constructed?
  – Poor man’s ensemble \textit{(distinct members)}
  – Multi-physics \textit{(distinct members)}
  – Random perturbation of initial conditions \textit{(anonymous members)}

• How are your forecasts used?
  – Improved point forecast \textit{(ensemble mean)}
  – Probability of an event
  – Full distribution
Approaches to evaluating ensemble forecasts

• As individual members
  – Use methods for continuous or categorical forecasts

• As probability forecasts
  – Create probabilities by applying thresholds or statistical post-processing

• As a full distribution
  – Use individual members or fit a distributions through post-processing
Evaluate each member as a separate, deterministic forecast

• Why? Because it is easy and important
  • If members are unique, it might provide useful diagnostics.
  • If members are biased, verification statistics might be skewed.
  • If members have different levels of bias, should you calibrate?

– Do these results conform to your understanding of how the ensemble members were created?
Verifying a probabilistic forecast

- You cannot verify a probabilistic forecast with a single observation.
- The more data you have for verification, (as with other statistics) the more certain you are.
- Rare events (low probability) require more data to verify.
- These comments refer to probabilistic forecasts developed by methods other than ensembles as well.
Properties of a perfect probabilistic forecast of a binary event.

Reliability

Resolution

Sharpness

- observed non-events
- observed events

Copyright UCAR 2015, all rights reserved.
The Brier Score

- Mean square error of a probability forecast

\[ BS = \frac{1}{n} \sum_{i=1}^{n} (f_i - x_i)^2 \]

where
- \( n \) is the number of forecasts
- \( f_i \) is the forecast prob on occasion \( i \)
- \( x_i \) is the observation (0 or 1) on occasion \( i \)

- Weights larger errors more than smaller ones
Brier Score

\[ BS = \frac{1}{n} \sum_{k=1}^{n} (f_k - x_k)^2 \]

where

\[ f_k = \text{forecast probability on occasion } k \]

\[ x_k = \text{observation (0 or 1) on occasion } k \]

BS can be decomposed into 3 components that represent important properties of the forecasts:

\[ BS = \frac{1}{n} \sum_{i=1}^{I} N_i (f_i - \bar{x}_i)^2 - \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{x}_i - \bar{x})^2 + \bar{x}(1 - \bar{x}) \]

Reliability  Resolution  Uncertainty

Where the \( I \) is the number of discrete values of \( f \) (e.g., \( f_1 = 0.05, f_2 = 0.10, f_3 = 0.20, \ldots \) etc.) and

\[ n = \sum_{i=1}^{I} N_i \]

\[ \bar{x}_i = \frac{1}{N_i} \sum_{k \in N_i} x_k \]

\[ \bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{1}{N} \sum_{i=1}^{I} N_i \bar{x}_i \]
Components of the Brier Score

• **Reliability**
  Measures how well the conditional relative frequency of events matches the forecast
  \[
  \frac{1}{n} \sum_{i=1}^{I} N_i (f_i - \overline{x}_i)^2
  \]

• **Resolution**
  Measures how well the forecasts distinguish situations with different frequencies of occurrence
  \[
  \frac{1}{n} \sum_{i=1}^{I} N_i (\overline{x}_i - \overline{x})^2
  \]

• **Uncertainty**
  Measures the variability in the observations (i.e., the difficulty of the forecast situations)
  \[
  \overline{x} (1 - \overline{x})
  \]

Looking at Brier Score components is critical to understand forecast performance
Brier Skill Score (BSS)

\[ \text{BSS} = \frac{\text{RES} - \text{REL}}{\text{UNC}} \]

BSS is a simple combination of the 3 components of the Brier Score (assumes “Sample Climatology” as the reference forecast)
Our friend, the scatterplot
Introducing the attribute diagram!

( close relative to the reliability diagram)

• Analogous to the scatter plot- same intuition holds.
• Data must be binned!
• Hides how much data is represented by each
• Expresses conditional probabilities.
• Confidence intervals can illustrate the problems with small sample sizes.
Attribute diagram shows reliability, resolution, skill.
Reliability Diagram Exercise
**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

---

(a) observed observed non-events events  
(b) observed observed non-events events  
(c) observed observed non-events events

<table>
<thead>
<tr>
<th>frequency</th>
<th>forecast</th>
<th>Good discrimination</th>
<th>Poor discrimination</th>
<th>Good discrimination</th>
</tr>
</thead>
</table>

Copyright UCAR 2015, all rights reserved.
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.
Interpretation of ROC

- 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.**

- ROC is **conditioned on the observations** (i.e., given that Y occurred, what was the corresponding forecast?)
- Reliability and ROC diagrams are good companions.

Copyright UCAR 2015, all rights reserved.
Relative Operating Characteristic (ROC)

ROC example:

ROC diagram for \(T12 < 5\) °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)
Sharpness also important

“Sharpness” measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable.

But: don’t want sharp if not reliable. Implies unrealistic confidence.
Sharpness ≠ resolution

• Sharpness is a property of the forecasts alone; a measure of sharpness in Brier score decomposition would be how populated the extreme $N_i$'s are.

$$BS = \frac{1}{n} \sum_{i=1}^{I} N_i (f_i - \bar{x}_i)^2 - \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{x}_i - \bar{x})^2 + \bar{x}(1 - \bar{x})$$
Sharpness for binary probability forecasts

For a binary probability forecast, sharpness is based on the distribution (histogram) of frequencies associated with each possible probability.

Sometimes summarized using the variance of the distribution.

Reasonable sharpness

Max possible sharpness

Perfect forecast

Poor sharpness
Forecasts of a full distribution

• How is it expressed?
  – Discretely by providing forecasts from all ensemble members
  – A parametric distribution – normal (ensemble mean, spread)
  – Smoothed function – kernel smoother
Evaluating ensembles

Rank Histograms

Ideal

Too wide

Too narrow

Continuous Ranked Probability Score:
Measures skill using squared error (analogous to MAE)
Ensemble Calibration / Reliability

• By default, we assume all ensemble forecasts have the same number of members. Comparing forecasts with different number of members is an advanced topic.

• For a perfect ensemble, the observation comes from the same distribution as the ensemble.
Rank histograms are a way to examine the calibration of an ensemble.

Copyright UCAR 2015, all rights reserved.
Creating rank histograms

Rank 1 of 21

Rank 14 of 21

Rank 5 of 21

Rank 3 of 21

Copyright UCAR 2015, all rights reserved.
Rank histograms are a way to examine the calibration of an ensemble.
Verifying a continuous expression of a distribution (i.e. normal, Poisson, beta)

• Probability of any observation occurring is on [0,1] interval.
• Probability Integral Transformed (PIT) - fancy word for how likely is a given forecast
• Still create a rank histogram using bins of probability of observed events.
Verifying a distribution forecast

Copyright UCAR 2015, all rights reserved.
Warnings about rank histograms

• Assume all samples come from the same climatology!

• A flat rank histogram can be derived by combining forecasts with offsetting biases


• Techniques exist for evaluating “flatness”, but they mostly require much data.
Continuous and discrete rank probability scores

• Measures of accuracy for
  – Multiple category forecasts (e.g., precipitation type)

  **Rank Probability Score (RPS)**
  – Continuous distributions (e.g., ensemble distribution)

  **Continuous Ranked Probability Score (CRPS)**

• Relates to Brier score – for a forecast of a binary event, the RPS score is equivalent to the Brier score.
Rank Probability Scores

(a) Forecast PDF and Observed

(b) Forecast and Observed CDF
A good RPS score minimizes area

(a) Forecast PDF and Observed

(b) Forecast and Observed CDF
Ignorance score (for multi-category or ensemble forecasts)

- A “local” score

\[ IS = \frac{1}{n} \sum_{i=1}^{n} \log_2 \left( \hat{p}_{t,k^*(t)} \right) \]

- \( k^*(t) \) is the category that actually was observed at time \( t \)
- Based on information theory
- Only rewards forecasts with some probability in “correct” category
- Perfect score: 0 [i.e., \( \log_2(1) = 0 \)]
Final comments

• Know how and why your ensemble is being created.
• Use a combination of graphics and scores.
• Areas of more research
  – Verification of spatial forecasts
  – Additional intuitive measures of performance for probability and ensemble forecasts.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Attribute evaluated</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Probability forecasts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brier score</td>
<td>Accuracy</td>
<td>Based on squared error</td>
</tr>
<tr>
<td>Resolution</td>
<td>Resolution (resolving different categories)</td>
<td>Compares forecast category climatologies to overall climatology</td>
</tr>
<tr>
<td>Reliability</td>
<td>Calibration</td>
<td></td>
</tr>
<tr>
<td>Skill score</td>
<td>Skill</td>
<td>Skill involves <em>comparison</em> of forecasts</td>
</tr>
<tr>
<td>Sharpness measure</td>
<td>Sharpness</td>
<td>Only considers distribution of forecasts</td>
</tr>
<tr>
<td>Relative Operating Characteristic (ROC)</td>
<td>Discrimination</td>
<td>Ignores calibration</td>
</tr>
<tr>
<td>C/L Value</td>
<td>Value</td>
<td>Ignores calibration</td>
</tr>
<tr>
<td><strong>Ensemble distribution</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank histogram</td>
<td>Calibration</td>
<td>Can be misleading</td>
</tr>
<tr>
<td>Spread-skill</td>
<td>Calibration</td>
<td>Difficult to achieve</td>
</tr>
<tr>
<td>CRPS</td>
<td>Accuracy</td>
<td>Squared difference between forecast and observed distributions Analogous to MAE in limit</td>
</tr>
<tr>
<td>log p score</td>
<td>Accuracy</td>
<td>Local score, rewards for correct category; infinite if observed category has 0 density</td>
</tr>
</tbody>
</table>
Useful references

- **Good overall references** for forecast verification:

- **Verification tutorial** – Eumetcal (http://www.eumetcal.org/-learning-modules-)


- **Brier score, continuous ranked probability score, reliability diagrams**: Wilks text again.


- **Economic value diagrams**:

Reliability Diagram Exercise

- Probabilities underforecast
- Essentially no skill
- Tends toward mean but has skill
- Small samples in some bins
- Reliable forecast of rare event
- No resolution
- Over-resolved forecast
- Typical categorical forecast