Ensemble Verification

Yuejian Zhu

Environmental Modeling Center
NOAA/NWS/NCEP

Acknowledgements:
Zoltan Toth EMC
Outlines

- Climatological Data
- Verify Analysis (proxy truth)
- RMS and Spread
- Mean Error and Absolute Error
- Histogram and Outlier
- RPS and RPSS
- CRPS and CRPSS
- BSS (Resolution and Reliability)
- ROC (Hit Rate and False Alarm Rate)
- Economic Value (cost-loss analysis)
Climatological Data

- NCEP/NCAR 40 years (1958-1997) reanalysis
- Monthly Sampling
  - For example: $40 \times 30 = 1200$
- 10 equally-a-likely, based on sampling
- Projected to verify date
- All forecast skills will base on 10 equally-a-likely climatological bins.
Verify Analysis (proxy truth)

• All following deterministic and probabilistic verification are based on 2.5*2.5 grid forecast, analysis and climatology in globally
• NCEP best analysis (GSI) is our best reference (proxy truth) to apply all NCEP forecast verifications.
• Other model forecast verification is using their own available analysis (proxy truth)
• For jointed ensemble (or multi-model ensemble), it is using NCEP analysis (as truth) in practice.
Due to model imperfection?
One day advantage
Prob. Evaluation (simple measurement)

1. Talagrand Distribution (histogram distribution):
   Sorting forecast in order, to check where the analysis is falling
   Reliability measurement, system bias detected.
   positive/negative biased for forecasting model, example of these forecasts --> cold bias,
   assume analysis is bias-free (perfect). Common -"U" sharp
Northern Hemisphere 500hPa Height Histogram Distribution

Average For 20061201 - 20070222

1-Day (%) | 3-Day (%)
---|---
10 | 10
5 | 5
0 | 0

5-Day (%) | 8-Day (%)
---|---
10 | 10
5 | 5
0 | 0

12-Day (%) | 16-Day (%)
---|---
10 | 10
5 | 5
0 | 0

Symbols:
- E14s
- E14sb
Prob. Evaluation (simple measurement)

1. Talagrand distribution (continue).
   . Outlier evolution by different leading time
   .. Adding up two outliers subtract the average.
   … Ideal forecasts will have zero outliers.

Due to inability of ensemble to capture model related errors?
**Prob. Evaluation** (simple measurement)

Outlier --> diagnostic forecasts vs. next forecasts (f+24hrs valid at same time)

Assume forecasting model is perfect, f+24.

Perfect forecast system will expect the outliers are zero.

---

**Detecting model initial uncertainty?**
**Prob. Evaluation** (multi-categories)

Based on climatological equally likely bins (for example, 5 bins)

For verifying multi-category probability forecasts.

measure both reliability and resolution.

1. Ranked (ordered) probability score (RPS) and RPSS

   \[ RPSS = \frac{RPS_f - RPS_c}{1 - RPS_c} \]

   \[ RPS (p, d) = 1 - \frac{1}{k-1} \left[ \sum_{i=1}^{k} \left( \sum_{n=1}^{i} P_n - \sum_{n=1}^{i} d_n \right)^2 \right] \]
Northern Hemisphere 500 mb Height
Ranked Probability Skill Scores (RPSS)
Yearly Average

What is THORPEX’s goal for 10 years?
Continuous Rank Probability Score

\[ CRPS = \int_{-\infty}^{+\infty} [F(x) - H(x - x_0)]^2 \, dx \]

\[ CRPSS = \frac{CRPS_c - CRPS_f}{CRPS_c} \]

Heaviside Function \( H \)

\[ H(x - x_0) = \begin{cases} 0 & (x \leq x_0) \\ 1 & (x > x_0) \end{cases} \]

Order of 10 ensemble members (p01, p02, ..., p10)
**Prob. Evaluation** (multi-categories)

2. Brier Score (BS, non-ranked), Brier Skill Score (BSS). From two categories to multi-categories/probabilistic ---- measure both reliability and resolution.
Prob. Evaluation (multi-categories)

3. Decomposition of Brier Score:
   consider **sub-sample** and **overall-sample**
   reliability, resolution and uncertainty.
   for reliability: 0 is perfectly reliable
   for resolution: 0 is no resolution ( = climatology )
   when resolution = reliability \(\rightarrow\) no skill

example of global ensemble:

![Graph](image)

Northern Hemisphere 500 mb Height Brier Skill Scores (BSS)
Average For 20020101 - 20020131

**No skill beyond this point**

**reliability**

**resolution**
Prob. Evaluation (multi-categories)

4. Reliability and possible calibration (remove bias):
For period precipitation evaluation

\[ BS = RELI - RESO + UNCE \]

\[ BSS = \frac{RESO - RELI}{UNCE} \]
Prob. Evaluation (multi-categories)

4. Reliability and possible probabilistic calibration:
re-label fcst prob by obs frequency associated with fcst

calibrated

Un-calibrated
**Prob. Evaluation** (cost-loss analysis)

Based on hit rate (HR) and false alarm (FA) rate.

1. Relative Operating Characteristics (ROC) area - *Appl. of signal detection theory for measuring discrimination between two alternative outcome.*

\[ \text{ROC area} = \text{Integrated area} \times 2 \quad (0-1 \text{ normality}) \]

<table>
<thead>
<tr>
<th>o/f</th>
<th>y(f)</th>
<th>n(f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y(o)</td>
<td>h</td>
<td>m</td>
</tr>
<tr>
<td>n(o)</td>
<td>f</td>
<td>c</td>
</tr>
</tbody>
</table>

**Relative Operating Characteristics**

ROC (Relative Operating Characteristics) curve for a 10-member T62 ensemble of forecasts and for T126 and T62 control forecasts for the 500 hPa height, *NH extratropics*, March–May 1997. The closer a curve is to the upper left hand corner, the more ability the forecasting system has in delineating between cases when a certain event (in this case, the occurrence of one of 10 climatologically equally likely bins) did or did not occur.
Relative Operating Characteristics area (ROC area)

- f(noise)
- f(signal)

Near perfect forecast

No skill forecast

Real forecast

Decision threshold

False alarm rate

Hit rate
USER NEEDS – PROBABILISTIC FORECAST INFORMATION FOR MAXIMUM ECONOMIC BENEFIT

ECONOMIC VALUE OF FORECASTS

Given a particular forecast, a user either does or does not take action (e.g., protects its crop against frost) *Mylne & Harrison, 1999*

**FORECAST**

<table>
<thead>
<tr>
<th>Observation</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>H(its)</td>
<td>M(isses)</td>
</tr>
<tr>
<td>Mitigated Loss</td>
<td></td>
<td>Loss</td>
</tr>
<tr>
<td>NO</td>
<td>F(alse alarms)</td>
<td>C(correct rejections)</td>
</tr>
<tr>
<td>Cost</td>
<td>No Cost</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Mean Expense}_{fc} = hML + mL + fC
\]

\[
\text{Mean Expense}_{perf} = oML
\]

\[
\text{Value} = \frac{\text{ME}_{cl} - \text{ME}_{fc}}{\text{ME}_{cl} - \text{ME}_{perf}}
\]

\[
\text{ME}_{cl} = \text{min}[oL, \text{oML} + (1-o)C]
\]

\[o=\text{climatological frequency}\]

Optimum decision criterion for user action: \( P(\text{weather event}) = C/L \)  
(*Murphy 1977*)
Prob. Evaluation (cost-loss analysis)

2. Economic Value (EV) of forecasts.

Given a particular forecast, a user either does or does not take action.
**Prob. Evaluation** (cost-loss analysis)
Based on hit rate (HR) and false alarm (FA) analysis
.. Economic Value (EV) of forecasts

![Graph showing skill scores over forecast days for deterministic and ensemble forecasts.](image)
## Decision Theory Example

### Critical Event: sfc winds > 50kt

- Cost (of protecting): $150K
- Loss (if damage): $1M

<table>
<thead>
<tr>
<th>Case</th>
<th>Deterministic Forecast (kt)</th>
<th>Observed (kt)</th>
<th>Cost ($K)</th>
<th>Probabilistic Forecast</th>
<th>Cost ($K) by Threshold for Protective Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
<td>54</td>
<td>150</td>
<td>42%</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>58</td>
<td>63</td>
<td>150</td>
<td>71%</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>73</td>
<td>57</td>
<td>150</td>
<td>95%</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>37</td>
<td>150</td>
<td>13%</td>
<td>150</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
<td>31</td>
<td>0</td>
<td>3%</td>
<td>150</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>55</td>
<td>1000</td>
<td>36%</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>71</td>
<td>150</td>
<td>85%</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>53</td>
<td>42</td>
<td>150</td>
<td>22%</td>
<td>150</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>27</td>
<td>0</td>
<td>51%</td>
<td>150</td>
</tr>
<tr>
<td>10</td>
<td>52</td>
<td>39</td>
<td>150</td>
<td>77%</td>
<td>150</td>
</tr>
</tbody>
</table>

**Total Cost:** $2,050

- **Hit:** $150K
- **False Alarm:** $150K
- **Miss:** $1000K
- **Correct Rejection:** $0K

**Optimal Threshold = 15%**