Evaluating Forecast Quality

Simon J. Mason

International Research Institute for Climate Prediction
Questions

• How do we decide whether a forecast was “correct”?  
• How do we decide whether a set of forecasts is correct consistently enough to be considered “good”?  
• How can we answer any of these questions when forecasts are expressed probabilistically?
How do we decide whether a forecast was “correct”?

<table>
<thead>
<tr>
<th></th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deterministic</strong></td>
<td>It will rain tomorrow</td>
<td>There will be 10 mm of rain tomorrow</td>
</tr>
<tr>
<td><strong>Probabilistic</strong></td>
<td>There is a 50% chance of rain tomorrow</td>
<td>There is a 50% chance of more than 10 mm of rain tomorrow</td>
</tr>
</tbody>
</table>
The method of forecast verification depends upon the type of information provided in the forecast:

**Deterministic - discrete**

Forecast: It will rain tomorrow.
Verification: Does it rain? Yes or no?

**Deterministic - continuous**

Forecast: There will be 10 mm of rain tomorrow.
Verification: Does it rain (approximately) 10 mm?
Probabilistic – discrete and continuous

Forecast: There is a 50% chance of (more than 10 mm of) rain tomorrow.

Verification: Yes!

As long as what happened was not given a 0% chance of occurring, a probabilistic forecast cannot be wrong.
Conclusion

• “Prediction is very difficult, especially if it’s about the future.”  
  Nils Bohr

• No – probabilistic forecasting is very easy because the forecast is never wrong!

?
How do we decide whether a set of forecasts is correct consistently enough to be considered “good”?

Deterministic - discrete

<table>
<thead>
<tr>
<th>OBSERVATIONS</th>
<th>FORECASTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>Hit</td>
</tr>
<tr>
<td>No</td>
<td>False alarm</td>
</tr>
</tbody>
</table>
**Hit Rate**

How many of the events were forecast?

\[
\text{Hit rate} = \frac{\text{number of hits}}{\text{number of events}} \times 100\%
\]

“Wall Street indices predicted nine out of the last five recessions!” *Newsweek*, 19 Sept., 1966.

**Correct rejections** are correct forecasts.

\[
\text{False-alarm rate} = \frac{\text{number of false alarms}}{\text{number of nonevents}}
\]
Hit Score

How many times was the forecast correct?

Hit score = \( \frac{\text{number of correct forecasts}}{\text{number of forecasts}} \times 100\% \)

Hit score = \( \frac{\text{number of hits and correct rejections}}{\text{number of forecasts}} \times 100\% \)

A measure of forecast accuracy.
Finley’s Tornado Forecasts

A set of tornado forecasts for the U.S. Midwest published in 1884.

<table>
<thead>
<tr>
<th>OBSERVATIONS</th>
<th>FORECASTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tornado</td>
<td>No tornado</td>
</tr>
<tr>
<td>Tornado</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>No tornado</td>
<td>72</td>
<td>2680</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>2703</td>
</tr>
</tbody>
</table>

Hit score \( \text{Hit score} = \frac{28 + 2680}{2803} \times 100\% = 96.6\% \)
No Tornado Forecasts

A better score can be achieved by issuing no forecasts of tornadoes!

<table>
<thead>
<tr>
<th>OBSERVATIONS</th>
<th>FORECASTS</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tornado</td>
<td>No tornado</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Tornado</td>
<td>0</td>
<td>51</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>No tornado</td>
<td>0</td>
<td>2752</td>
<td>2752</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>2803</td>
<td>2803</td>
<td></td>
</tr>
</tbody>
</table>

Hit score = \( \frac{0 + 2752}{2803} \times 100\% = 98.2\% \)
Problems

**Accuracy**: The forecasts correspond well with the observations, but:

**Uncertainty**: Tornadoes are rare events so it is easy to score a correct rejection. There is very little uncertainty.

**Skill**: The accuracy of one set of forecasts compared to that of another set.
How do we decide whether a set of forecasts is “good”?

Is this set of forecasts better than another (simple) set?

Often the comparison is with a simple forecast strategy such as:

- climatology;
- persistence;
- perpetual forecasts;
- random forecasts.
Hit Skill Score

How many more times was the forecast correct compared to a reference forecast strategy?

\[
\text{Hit skill score} = \frac{\# \text{ correct} - \# \text{ expected correct}}{\# \text{ forecasts} - \# \text{ expected correct}} \times 100\%
\]

\[
\text{Hit skill score} = \frac{(28 + 2680) - 2752}{2803 - 2752} \times 100\% = -86.3\%
\]

Exercise: calculate the hit skill score for Finley’s forecasts compared to a strategy of random guessing.
How do we decide whether a set of forecasts is correct consistently enough to be considered “good”?

**Deterministic – continuous**

**Correlation**

A measure of *association* – e.g., does rainfall increase if forecast rainfall increases?
A set of “perfect” forecasts!
Mean squared error

\[
\text{mean squared error} = \frac{\text{total of squared errors}}{\text{number of forecasts}}
\]

A measure of forecast accuracy.

Composed of:

- correlation;
- mean bias;
- amplitude bias.

Can be converted to a skill score.
How do we decide whether one set of probabilistic forecasts is better than another?

**Brier score**

Measures the mean-squared error of probability forecasts.

\[
\text{Brier score} = \frac{\text{total of squared probability errors}}{\text{number of forecasts}}
\]

If an event was forecast with a probability of 60%, and the event occurred, the probability error is:

\[
60\% - 100\% = -40\%
\]
Verification of probabilistic forecasts

- How do we know if a probabilistic forecast was “correct”?

  “A probabilistic forecast can never be wrong!”

As soon as a forecast is expressed probabilistically, all possible outcomes are forecast. However, the forecaster’s level of confidence can be “correct” or “incorrect” = reliable.

Is the forecaster over- / under-confident?
Verification of probabilistic forecasts

• How do we know if a forecaster is over- / under-confident

Whenever a forecaster says there is a high probability of rain tomorrow, it should rain more frequently than when the forecaster says there is a low probability of rain.
Forecast reliability

- A forecast is **consistent** if the forecast probability is a true estimate of the forecaster’s level of confidence.
- If forecasts are **reliable**, the forecaster’s confidence is appropriate.
- If forecasts are **consistent** and **reliable**, the probability that the event will occur is the same as the forecast probability.
Desired characteristics of forecasts:

Probabilities should be **reliable**.
Probabilities should be **sharp**.

**Reliability** is a function of forecast **accuracy**.

Assuming the forecasts are **reliable**, **sharpness** is a function of predictability.
For all forecasts of a given confidence, identify how often the event occurs. If the proportion of times that the event occurs is the same as the forecast probability, the probabilities are reliable (or well calibrated).

A plot of relative frequency of occurrence against forecast probability will be a diagonal line if the forecasts are reliable.

Problem: large number of forecasts required.
Reliability diagrams for forecasts of Nino3.4 sea-surface temperatures.
Reliability Diagrams

Exercise: diagnose the following reliability curves.
Relative (or Receiver) Operating Characteristics (ROC)

Convert probabilistic forecasts to deterministic forecasts by issuing a warning if the probability exceeds a threshold minimum.

By raising the threshold less warnings are likely to be issued - reducing the potential of issuing a false alarm, but increasing the potential of a miss.

By lowering the threshold more warnings are likely to be issued - reducing the potential of a miss, but increasing the potential of a false alarm.
Relative (or Receiver) Operating Characteristic (ROC)

Are (proportionately) more warnings issued for events than for non-events?

Plot the hit rate against the false alarm rate for varying thresholds.

The ROC is used to estimate whether forecasts are potentially useful.
Hit rate

Defines: the proportion of events for which a warning was provided correctly

Estimates: the probability that an event will be forewarned

hit rate = \frac{\text{number of hits}}{\text{number of events}}
False alarm rate

Defines: the proportion of non-events for which a warning was provided incorrectly

Estimates: the probability that a warning will be provided incorrectly for a non-event

\[
\text{false alarm rate} = \frac{\text{number of false alarms}}{\text{number of nonevents}}
\]
Relative (or Receiver) Operating Characteristics (ROC)

ROC curves for ECHAM 3.6 simulations of March-May below- and above-normal precipitation over eastern Africa (10°-10°S, 30°-50°E)
Conclusions

• Even for deterministic forecasts, there is no single measure that gives a comprehensive summary of forecast quality:
  - accuracy
  - skill
  - uncertainty
• Probabilistic forecasts address the two fundamental questions:
  - What is going to happen?
  - How confident can we be that it is going to happen?
• Both these aspects require verification.
Conclusions

• The most important aspects of forecast quality:
  - the most likely outcome must be the one that occurs most frequently;
  - confidence in the forecast must be appropriate (reliability);
  - forecast probabilities should be sharp (without compromising reliability).

• Good forecasts must address:
  - consistency
  - quality
  - value
Recommended readings

Forecast verification and quality: