Section 5: Forecast Evaluation and Skill Scores
What is Forecast Evaluation?

- Assessing the quality / error structure of forecasts by comparison to independent observations

Model

Input / Conditions

Forecast: Statement about Reality

Reality / Observations

Skill scores:
Measures of forecast quality
“Forecasts”

• **Weather Forecast**
  How accurate are temperature forecasts one day ahead?

• **Simulations of Climate**
  Reproduce the distribution of mean summer precipitation in Europe?

• **Spatial analysis**
  Estimate precipitation at a non-instrumented site from observations in the neighbourhood?

• **Remote sensing, …**
Observations

• **Generic for “measure of reality”**

• **The chosen Reference**

• **In practice:**
  - In-situ measurements
  - Indirect estimates of “reality”: re-analyses, remote sensing

• **Important:**
  - Role of observation errors for your evaluation?
  - Are observations and model independent?
Why Forecast Evaluation?

• Learn how to properly use / interpret forecast
  o E.g. the issuing of a public flood warning depends on the frequency with which the forecast produces false alarms

• Learn how and where to improve forecast
  o E.g. by comparison of forecast quality for different model parametrizations

• Justify investments made into models, instruments
  o E.g. launching of new weather satellites depends on the expected improvement of weather forecasts (pay-back on investment)
ECMWF MR-Forecast

Anomaly correlation of 500 hPa Geopotential

ECMWF 2012
Forecasts

- **Continuous:**
  - real value, e.g. temperature in Zürich

- **Categorial:**
  - values in discrete classes (e.g. cold, normal or warm) or events (e.g. a tornado tomorrow).

- **Deterministic:**
  - a single number, e.g. the expected temperature tomorrow

- **Probabilistic:**
  - probabilities, e.g. the prob. of rain tomorrow
  - expresses the degree of forecast uncertainty
Outline

• Deterministic categorial forecasts
• Deterministic continuous forecasts
• Probability forecasts
• Evaluation based on economic value

• Material based on:
  o Richardson 2000, Wilks 2001
Section 5: Forecast Evaluation and Skill Scores

Deterministic Categorial Forecasts
Contingency Table

- **Binary Forecasts**
  - $Y = \{\text{yes, no}\}$, e.g. events: tomorrow it will (will not) rain
  - simplest categorial case

- **Contingency Table**
  - Distribution $(Y,O)$

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observation</th>
<th>Marginal of Fcst</th>
<th>Marginal of Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>$a$ hits</td>
<td>$a+c$ yes obs</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>$b$ false alarms</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>$c$ misses</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>$d$ correct rejects</td>
<td>$b+d$ no obs</td>
</tr>
</tbody>
</table>

$N$ total fcsts

Obs. evts $\rightarrow$ fcst. evts
## Finley Tornado Forecasts 1884

U.S. Army forecasts of tornado occurrence east of the Rockies, based on synoptic information

<table>
<thead>
<tr>
<th>Tornados forecasted</th>
<th>Tornados Observed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>72</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>2752</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2803</td>
</tr>
</tbody>
</table>
Simple Scores

- **Bias score:**

\[
B = \frac{a + b}{a + c} = \frac{\text{forecasted events}}{\text{observed events}}
\]

- \( B = 1 \) unbiased, \( B < 1 \) underforecast, \( B > 1 \) overforecast
- depends on marginals only, does not measure ‘correspondence’

- **Probability of detection (hit rate):**

\[
POD = \frac{a}{a + c} = \frac{\text{hits}}{\text{observed events}}
\]

- Fraction of all observed events correctly forecasted
- \( 0 \leq POD \leq 1 \), best score: \( POD = 1 \), best score ≠ perfect fcst
- Focus on events. No penalty for false alarms.
Simple Scores

• False alarm ratio:

\[ FAR = \frac{b}{a+b} = \frac{\text{false alarms}}{\text{forecasted events}} \]

  o Fraction of forecasted events that were false alarms
  o \( 0 \leq FAR \leq 1 \), best score: \( FAR = 0 \), best score ≠ perfect fcst

• Probability of false detection (false alarm rate):

\[ POFD = \frac{b}{b+d} = \frac{\text{false alarms}}{\text{non-events}} \]

  o Fraction of all non-events when forecast predicted an event
  o \( 0 \leq POFD \leq 1 \), best score: \( POFD = 0 \), best score ≠ perfect fc
Simple Scores

- **Accuracy (fraction correct):**

\[
ACC = \frac{a + d}{N} = \frac{\text{correct forecasts}}{\text{all forecasts}}
\]

- Fraction of all forecasts that were correct
- \(0 \leq ACC \leq 1\), best score: \(ACC = 1\), best score = perfect fcst
- Events and non-events treated symmetrically
- For rare events the score is dominated by non-events

- Finley tornado forecast:
  - \(ACC = (28 + 2680)/2803 = 0.96\) (!)
  - But: \(POD = 28/51 = 0.54\) and \(FAR = 0.72\) (!)
Simple Scores

- Threat score (Critical Success Index):

\[ TS = CSI = \frac{a}{a + b + c} = \frac{\text{hits}}{\text{all forecasted or observed events}} \]

- Fraction of all forecasted or observed events that were correct
- \( 0 \leq TS \leq 1 \), best score: \( TS = 1 \), best score = perfect fcst
- Asymmetric between events and non-events.

- Finley tornado forecast:
  - \( TS = \frac{28}{(28+72+23)} = 0.23 \)
Limitations of Simple Scores

• How large is a “good” score?

• Best score not necessarily perfect forecast!

• Hedging (“Playing”) a score:
  o Example: Modify Finley’s Forecast --> constant forecast

<table>
<thead>
<tr>
<th>Forecasted</th>
<th>Observed</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>51</td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>2752</td>
<td></td>
</tr>
</tbody>
</table>

Finley: $ACC = 0.96$
Constant: $ACC = 0.98$ (!)
Generic Form of a Skill Score

\[ SS = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \]

- \( A \) accuracy score, e.g. \( ACC \) or \( TS \)
- \( A_{ref} \) accuracy of reference forecast, e.g. random
- \( A_{perf} \) accuracy of perfect forecast

- \( SS = 1 \) perfect forecast
- \( SS > 0 \) skillful, better than reference
- \( SS < 0 \) less skillful than reference
Heidke Skill Score

- Generic Score with …
  … ACC as A and random forecast as reference

\[
A = \left( \frac{a + d}{N} \right) \quad A_{perf} = 1
\]

\[
A_{ref} = \left( \frac{a + b}{N} \right) \cdot \left( \frac{a + c}{N} \right) + \left( \frac{d + c}{N} \right) \cdot \left( \frac{d + b}{N} \right)
\]

- Heidke Skill Score

\[
HSS = \frac{ad - bc}{((a + c) \cdot (c + d) + (a + b) \cdot (b + d))/2}
\]

\(-\infty < HSS \leq 1, \quad HSS \leq 0 \text{ no skill}\)
HSS for Finley Forecast

- **HSS**
  - for Finley forecast: $HSS=0.355$
  - for constant forecast: $HSS=0.0$

  - note, ACC is large even for random forecast:

  $$ACC_{random} = \left( \frac{28 + 72}{2803} \right) \cdot \left( \frac{28 + 23}{2803} \right) + \left( \frac{2680 + 23}{2803} \right) \cdot \left( \frac{2680 + 72}{2803} \right) = 0.947$$

- **HSS (generic form of skill scores) compensates for high random ACC, when events are very rare.**
Hanssen-Kuipers Discriminant

- Similar to HSS but unbiased ACC in denominator

\[
SS = \frac{ACC - ACC_{\text{random}}}{1 - ACC_{\text{unbiased random}}}
\]

\[
ACC_{\text{unbiased random}} = \frac{(a + c)^2 + (b + d)^2}{N^2}
\]

- Hanssen-Kuipers (also True Skill Statistic, Pierce Skill Score)

\[
HK = \frac{ad - bc}{(a + c) \cdot (b + d)} = POD - POFD
\]

- \(-1 \leq HK \leq 1, \ HK \leq 0\) no skill,
- for unbiased forecasts: \(HK = HSS\)
- \(HK(\text{Finley}) = 0.523, \ HK(\text{constant}) = 0.0\)
Example

Hanssen-Kuipers Score (in %)
for daily precipitation occurrence (P>1 mm)

LokalModell: Operational NWP model of DWD in 2002, dx = 7 km)

Evaluation for all grid points in Germany for year 2002

Skill varies between seasons:
E.g. 24h fcst in summer is less accurate than 48h fcst in winter.

U. Damrath (DWD)
Equitable Threat Score

- Equitable Threat Score (also Gilbert Skill Score)
  - Use $TS$ (CSI) for $A$ in generic form, random forecast as reference

$$ETS = \frac{a/(a+b+c) - a_{ref}/(a+b+c)}{1-a_{ref}/(a+b+c)} = \frac{a - a_{ref}}{a - a_{ref} + b + c}$$

$$a_{ref} = (a + c) \cdot (a + b)/N$$

- $-1/3 \leq ETS \leq 1$, $ETS \leq 0$ no skill,
- $ETS$(Finley) = 0.216, $ETS$(constant) = 0
- Unlike with $HSS$ and $HK$, with $ETS$ focus is on events only
Skill Scores Differ …

- ... in the relative importance of systematic and random errors
  - E.g. artificially biasing a forecast decreases $HK$ linearly but less than linearly for $HSS$

- ... in the relative role of events and non-events
  - $ETS$ values only events $\leftrightarrow HSS$, $HK$ value both

- ... in their behaviour for rare events
  - Most skill scores tend to approach 0 for more and more rare events

- There is no single best recommendation!
Uncertainty in Scores

- You’ve got 30 event forecasts. You obtain HSS=0.2. Not too bad but …

- … what is the probability that such a score is obtained by chance?
Further Remarks

• **Sampling uncertainty**
  - Accuracy of skill scores decreases with sample size
  - Scores for forecasts of very rare events may be difficult to determine accurately.
  - Use resampling methods to quantify skill uncertainty.

• **Multi-category skill scores:**
  - 2x2 Table --> $k \times k$ Table
  - Extend classical scores to multi-category case.
  - E.g. $ACC$ is sum of diagonal table elements divided by total forecasts.
  - Ordered multi-category case: introduce weights to penalize for elements more far off the diagonal. (Gerrity 1992, see Wilks p. 274)
Section 5: Forecast Evaluation and Skill Scores

Deterministic Continuous Forecasts
Notation

- Sample, forecast-observation pairs (real valued)

\[ \{ y_i, o_i \}, \quad i = 1 .. N \]

- Sample means

\[ \bar{y} = \frac{1}{N} \sum_i y_i, \quad \bar{o} = \frac{1}{N} \sum_i o_i \]

- Sample variance

\[ s_y^2 = \frac{1}{N} \sum_i (y_i - \bar{y})^2, \quad s_o^2 = \frac{1}{N} \sum_i (o_i - \bar{o})^2 \]
Example Data

- 24-h forecasts of T-max
  Oklahoma City

- Comparison of:
  - NWS: Human forecast
  - NGM, LFM: Numerical model forecasts with MOS
  - PER: Persistence forecast

- Here
  - 2 summers (1993/4, N=182)

Brooks & Doswell 1996
Simple Error Scores

- **Bias (mean error, systematic error):**
  - additive, multiplicative

  \[ B_{add} = \bar{y} - \bar{o}, \quad B_{mult} = \bar{y} / \bar{o} \]

- **Mean absolute error:**
  - Mean of absolute deviations from obs

  \[ MAE = \frac{1}{N} \sum_{i} |y_i - o_i| \]

- **Mean squared error (MSE), root MSE (RMSE):**

  \[ MSE = \frac{1}{N} \sum_{i} (y_i - o_i)^2, \quad RMSE = \sqrt{MSE} \]

  - Sensitive to outliers, dominated by large deviations
  - Favors forecasts avoiding large deviations from the mean
Simple Error Scores

- Root means squared fraction (RMSF):

\[ RMSF = \exp\left(\sqrt{\frac{1}{N} \sum_i \log\left(\frac{y_i}{o_i}\right)^2}\right) \]

- Similar to RMSE but for multiplicative errors
- “Average multiplicative error”
- Meaningful for rainfall, wind speed, visibility, … (>0 !)
- Log insures that multiplicative under- / overestimates are equally penalized.
- Perfect forecast: RMSF = 1

Golding 1998
Correlation Skill Score

- Linear correlation coeff.

\[ \rho = \frac{1}{N} \sum_{i}^{N} (y_i - \bar{y}) \cdot (o_i - \bar{o}) \]

- \(-1 \leq \rho \leq 1\), \(\rho = 1\) best score
- A measure of random error (scatter around best fit)
- Insensitive to biases and errors in variance
- \(\rho^2\): fraction of variance in obs explained by “best” linear model
- \(\rho\) measures potential skill (see also later)

Oklahoma JJA Daily Max. Temperatures

BIAS: -1.275
MAE: 1.86
RMSE: 2.16

Linear Regression:

\[ o_i = \beta \cdot y_i + a + \varepsilon_i \]
Conditional Bias

- Linear regression slope

\[ \beta = \frac{s_o}{s_y} \cdot \rho \]

- \( \beta = 1 \) best score
- Deviations of \( \beta \) from 1 measure conditional bias
- \( \beta > 1 \): Large (small) values tend to be under- (over-) estimated (unless compensated by absolute bias).
- \( \beta \) is a function of correlation and fraction of variances

Linear Regression:

\[ o_i = \beta \cdot y_i + a + \epsilon_i \]
Decomposition of RMSE

- **RMSE’ (debiased RMSE)**

\[
RMSE'^2 = (\bar{y} - \bar{d})^2 + s_y^2 + s_o^2 - 2s_y s_o \rho
\]

\[
\Rightarrow \quad \frac{RMSE'^2}{s_o^2} = \frac{RMSE^2 - B^2}{s_o^2} = 1 + \frac{s_y^2}{s_o^2} - 2 \frac{s_y}{s_o} \rho
\]

Relative error in variance

Degree of correspondence

- **Geometric interpretation (cosine triangle theorem):**

\[
\cos \kappa = \rho
\]

\[
RMSE' \quad / \quad s_o
\]

Taylor 2001
Derivation

\[ \text{RMSE}^2 = \frac{1}{N} \sum (y_i - o_i)^2 = \frac{1}{N} \sum ((y_i - \bar{y}) - (o_i - \bar{o}) + (\bar{y} - \bar{o}))^2 \]

\[ = \frac{1}{N} \sum ((y_i - \bar{y}) - (o_i - \bar{o}))^2 + \frac{1}{N} \sum (\bar{y} - \bar{o})^2 \]

\[ = s_y^2 + s_o^2 - 2s_y s_o \rho + B^2 \]

\[ \text{RMSE}^2 - B^2 = s_y^2 + s_o^2 - 2s_y s_o \rho \]
Taylor Diagram

• Visualisation of forecast performance by three related scores in one graph.

• Ideal for:
  o Comparing several forecast models,
  o Comparing to a reference forecast
  o Comparing to several observation datasets.
  o Assessing skill uncertainty e.g. by ensembles.

Oklahoma JJA Daily Max. Temperatures

Taylor 2001
Taylor Diagram

- **Visualisation of forecast performance by three related scores in one graph.**

- **Ideal for:**
  - Comparing several forecast models,
  - Comparing to a reference forecast
  - Comparing to several observation datasets.
  - Assessing skill uncertainty e.g. by ensembles.

![Taylor Diagram](image)

- NWS: human forecaster
- NGM, LFM: numerical models
- PER: persistence forecast

Taylor 2001
Quiz

• How will the points change with another obs. reference?

Indian Monsoon in global climate models (AMIP Models) (from Taylor 2001)
Reduction of Variance

\[
SS = \frac{MSE - MSE_{clim}}{MSE_{perfect} - MSE_{clim}} = 1 - \frac{MSE}{MSE_{clim}} = 1 - \frac{1}{N} \sum \frac{(y_i - o_i)^2}{s_o^2}
\]

- also called *Brier score* or *Nash-Sutcliffe Efficiency* (Hydrology)
- generic form of skill score with \( A = MSE \) and climatological forecast as reference.
- value range: \(-\infty < SS \leq 1\)
- perfect forecast: \( SS = 1 \)
- climatology forecast: \( SS = 0 \)
- random forecast with same variance and mean like observations: \( SS = -1 \)
- sensitive to biases and errors in variance
- Always: \( SS \leq \rho^2 \) (see later)
- Oklahoma Temperature Forecast (NGM): \( SS = 0.607 \) (\( \rho^2 = 0.77 \))
Murphy-Epstein Decomposition

- **Decomposition of \( SS \) (Reduction of Variance)**

\[
\frac{MSE}{MSE_{clim}} = \frac{RMSE^2}{s_o^2} = \frac{(\bar{y} - \bar{o})^2}{s_o^2} + 1 + \frac{s_y^2}{s_o^2} - 2 \frac{s_y}{s_o} \rho \\
\left( \frac{\rho - s_y}{s_o} \right)^2 - \rho^2
\]

\[
\Rightarrow \quad SS = 1 - \frac{MSE}{MSE_{clim}} = \rho^2 - \left( \frac{\rho - s_y}{s_o} \right)^2 - \frac{(\bar{y} - \bar{o})^2}{s_o^2}
\]

- Linear correspondence
  - “maximum explained variance”
- Penalty for **absolute bias**
- Penalty for **conditional bias**

Murphy & Epstein 1989
Murphy-Epstein Decomposition

• Implications
  ○ $SS = \rho^2$ only for absolute and conditionally unbiased forecasts. I.e. $\rho^2$ is a measure of potential skill.

  ○ A non-perfect forecast ($\rho^2 < 1$) can only be conditionally unbiased if $s_y < s_o$, i.e. if variance is underestimated.

  ○ Conditional bias can be minimized by setting $s_y/s_o = \rho$, i.e. $SS$ can be “played”!

  ○ Among forecasts with the same $\rho$ and the same absolute bias, $SS$ (and $RMSE$) favors those with small conditional bias, i.e. too smooth forecasts.

  ○ Forecasts with “good variance” are generally handicaped.
# Oklahoma Temperatures

<table>
<thead>
<tr>
<th>Model</th>
<th>$\rho^2$</th>
<th>(Conditional bias)$^2$</th>
<th>(Absolute bias)$^2$</th>
<th>$SS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWS</td>
<td>0.824</td>
<td>0.002</td>
<td>0.000</td>
<td>0.822</td>
</tr>
<tr>
<td>NGM</td>
<td>0.771</td>
<td>0.026</td>
<td>0.138</td>
<td>0.607</td>
</tr>
<tr>
<td>LFM</td>
<td>0.750</td>
<td>0.002</td>
<td>0.000</td>
<td>0.748</td>
</tr>
<tr>
<td>PER</td>
<td>0.382</td>
<td>0.141</td>
<td>0.000</td>
<td>0.241</td>
</tr>
</tbody>
</table>

$\beta < 1$, because $s_y = s_o$
Summary

• Correlation is a measure of potential skill only.

• A thorough assessment of forecast quality requires consideration of several skill scores.

• Frequently used scores favor smooth forecasts. It is difficult to demonstrate skill of high variability forecasts.

• Use creative graphics (such as the Taylor diagram) to visualize several skill measures.