Measures of Skill

Simple Methods for Describing Accuracy, Inaccuracy, and others

With emphasis on different characteristics
Wilks' chapter 7.3.3 and 7.4 plus additional sources
http://www.cawcr.gov.au/projects/verification/verif_web_page.html

Mean Square Errors

- Often, the root mean square error is given to put the diagnostic in the same units as the original data.

$$MSE_{\text{clim}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

- Where this is the MSE of data (x_i) compared to a climatological mean for a stationary (non-changing) climatology.

- For persistence, the MSE can be calculated as a difference from persistence

$$MSE_{\text{Persistence}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - x_{i-1}\right)^2$$

MSE with a non-Stationary Climatology

- If the climatology is changing (e.g., daily temperatures changing over a season or two), then we don't to consider this variability in the climatology in the measure of error.
- A measure of error might be the MSE minus the MSE for the climatology:

$$UNEXPLAINED\ MSE = MSE - MSE_{\text{clim}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i} - \overline{x}\right)^{2} - \frac{1}{n}\sum_{i=1}^{n}\left(x_{CLIM,\,i} - \overline{x}\right)^{2}$$

Skill Scores

- In general, skill scores (*SS*) are defined as

$$SS = \frac{MSE - MSE_{\text{clim}}}{0 - MSE_{\text{clim}}} = 1 - \frac{MSE}{MSE_{\text{clim}}}$$

- A measure of skill (commonly called a skill score, SS) might be the ratio of the unexplained MSE to the climatological MSE:

$$UNEXPLAINED\ \ MSE = MSE - MSE_{\text{clim}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i} - \overline{x}\right)^{2} - \frac{1}{n}\sum_{i=1}^{n}\left(x_{CLIM\,,\,i} - \overline{x}\right)^{2}$$

- Where *A* is a measure of accuracy (or inaccuracy)

$$SS = \frac{A - A_{\text{ref}}}{A_{perfect} - A_{\text{ref}}} x100\%$$

- This *A* can be any metric. The SS will be the skill in this metric.

Different Wording

- Wording for forecasts (right) vs. hypothesis testing (left)

Reality About Null Hypothesis	Statistical Inference About Null Hypothesis: Accepted	Statistical Inference About Null Hypothesis: Rejected
True	Correct Acceptance	False Rejection (α)
False	False Acceptance (β)	Correct Rejection

Forecast	Event Observed? Yes	Event Observed? No
occurred	Hit	False Alarm
Did not occur	Miss	Correct Negative

- In this application, we typically take the Reality true or false to be either zero or one.
- In this case, event either occur or don't occur.

Probability of Detection (POD)

- *What fraction of the observed "yes" events were correctly forecast?*
- **Range:** 0 to 1.
- **Perfect score:** 1.
- **Characteristics:** Sensitive to hits, but ignores false alarms.
- Very sensitive to the climatological frequency of the event. Good for rare events.
- Can be artificially improved by issuing more "yes" forecasts to increase the number of hits. Should be used in conjunction with the false alarm ratio (below).
- *POD* is also an important component of the [Relative Operating Characteristic (ROC)] used widely for probabilistic forecasts.

Event Observed?

Forecast	Yes	No
Correct	Hit	False Alarm
Wrong	Miss	Correct Negative

$$POD = \frac{hits}{hits + misses}$$

Probability of Detection Example 1

Forecast	Event Observed? Yes	Event Observed? No
Correct	Hit	False Alarm
Wrong	Miss	Correct Negative

Forecast	Event Observed? Yes	Event Observed? No
Correct	120	10
Wrong	20	300

$$POD = \frac{hits}{hits + misses}$$

$$POD = \frac{120}{120 + 20} = 85.7\%$$

Probability of Detection Example 2

		Event Observed?	
		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

		Event Observed?	
		Yes	No
Forecast	Correct	120	3000
	Wrong	20	300

$$POD = \frac{hits}{hits + misses}$$

$$POD = \frac{120}{120+20} = 85.7\%$$

- POD is completely insensitive to False Alarms
- Potentially very misleading if limitations are not known!

False Alarm Rate or Ratio (FAR)

- ***Answers the question:** What fraction of the predicted "yes" events actually did not occur (i.e., were false alarms)?*
- **Range:** 0 to 1.
- **Perfect score:** 0.
- **Characteristics:** Sensitive to false alarms, but ignores misses.
- Very sensitive to the climatological frequency of the event.
- Should be used in conjunction with the probability of detection.

		Event Observed?	
		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

False Alarm Rate or Ratio (FAR) Example 1

		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

		Yes	No
Forecast	Correct	120	10
	Wrong	20	300

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

$$FAR = \frac{10}{120+10} = 7.7\%$$

➢

False Alarm Rate or Ratio (FAR) Example 2

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Forecast		Yes	No
	Correct	120	3000
	Wrong	20	300

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

$$FAR = \frac{3000}{120 + 3000} = 96\%$$

➢

Probability of False Detection (POFD)

- ***Answers the question:*** *What fraction of the observed "no" events were incorrectly forecast as "yes"?*
- **Range:** 0 to 1.
- **Perfect score:** 0.
- **Characteristics:** Sensitive to false alarms, but ignores misses.
- Can be artificially improved by issuing fewer "yes" forecasts to reduce the number of false alarms.
- Not often reported for deterministic forecasts, but is an important component of the [Relative Operating Characteristic (ROC)]() used widely for probabilistic forecasts.

Event Observed?

Forecast	Yes	No
Correct	Hit	False Alarm
Wrong	Miss	Correct Negative

$$PODF = \frac{false\ alarms}{correct\ \ negatives + false\ alarms}$$

Probability of False Detection (POFD) Example 1

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

Forecast		Yes	No
	Correct	120	10
	Wrong	20	300

$$PODF = \frac{false\ alarms}{correct\ negatives + false\ alarms}$$

$$PODF = \frac{10}{300+10} = 3.2\%$$

➢

Probability of False Detection (POFD) Example 2

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

Forecast		Yes	No
	Correct	120	3000
	Wrong	20	300

$$PODF = \frac{false\ alarms}{correct\ negatives + false\ alarms}$$

$$PODF = \frac{3000}{300 + 3000} = 91\%$$

➢

Threat Score or Critical Success Index (CSI)

- ***Answers the question:*** *How well did the forecast "yes" events correspond to the observed "yes" events?*
- **Range:** 0 to 1, 0 indicates no skill.
- **Perfect score:** 1.
- **Characteristics:** Measures the fraction of observed and/or forecast events that were correctly predicted.
- It can be thought of as the *accuracy* when correct negatives have been removed from consideration
  - *TS* is only concerned with forecasts that count.
- Sensitive to hits, penalizes both misses and false alarms.
- Does not distinguish source of forecast error.
- Depends on climatological frequency of events (poorer scores for rarer events) since some hits can occur purely due to random chance.

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

$$CSI = \frac{hits}{hits + misses + \ false\ alarms}$$

Critical Success Index (CSI): Example 1

		Event Observed? Yes	Event Observed? No
Forecast	Correct	Hit	False Alarm
Forecast	Wrong	Miss	Correct Negative

		Event Observed? Yes	Event Observed? No
Forecast	Correct	120	10
Forecast	Wrong	20	300

$$CSI = \frac{hits}{hits + misses + false\ alarms}$$

$$CSI = \frac{120}{120 + 20 + 10} = 80\%$$

➢

Critical Success Index (CSI): Example 2

		Event Observed? Yes	Event Observed? No
Forecast	Correct	Hit	False Alarm
Forecast	Wrong	Miss	Correct Negative

		Event Observed? Yes	Event Observed? No
Forecast	Correct	120	3000
Forecast	Wrong	20	300

$$CSI = \frac{hits}{hits + misses + false\ alarms}$$

$$CSI = \frac{120}{120 + 20 + 3000} = 3.8\%$$

➢

Equitable Threat Score (ETS)

- ***Answers the question:*** *How well did the forecast "yes" events correspond to the observed "yes" events (accounting for hits due to chance)?*
- **Range:** -1/3 to 1, 0 indicates no skill.
- **Perfect score:** 1.
- **Characteristics:** Measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance.
- The *ETS* is often used in the verification of rainfall in NWP models because its "equitability" allows scores to be compared more fairly across different regimes.
- Sensitive to hits.
- Because it penalizes both misses and false alarms in the same way, it does not distinguish the source of forecast error.

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

$$ETS = \frac{hits - hits_{random}}{hits + misses + \ false\ alarms - hits_{random}}$$

$$hits_{random} = \frac{(hits + misses)(hits + \ false\ alarms)}{total}$$

Equitable Threat Score (ETS): Example 1

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

Forecast		Yes	No
	Correct	120	10
	Wrong	20	300

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

$$ETS = \frac{120 - 40.44}{120 + 20 + 10 - 40.44} = 72.6\%$$

$$hits_{random} = \frac{(hits + misses)(hits + false\ alarms)}{total}$$

$$hits_{random} = \frac{(120 + 20)(120 + 3000)}{120 + 20 + 10 + 300} = 40.44$$

Equitable Threat Score (ETS): Example 2

Event Observed?

		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

		Yes	No
Forecast	Correct	120	3000
	Wrong	20	300

$$ETS = \frac{hits - hits_{random}}{hits + misses + \ false\ alarms - hits_{random}}$$

$$ETS = \frac{120 - 127}{120 + 20 + 300 - 127} = -0.2\%$$

$$hits_{random} = \frac{(hits + misses)(hits + \ false\ alarms)}{total}$$

$$hits_{random} = \frac{(120 + 20)(120 + 3000)}{120 + 20 + 3000 + 300} = 127$$

Hanssen and Kuipers discriminant (HK)

- ***Answers the question:*** *How well did the forecast separate the "yes" events from the "no" events?*
- **Range:** -1 to 1, 0 indicates no skill.
- **Perfect score:** 1.
- **Characteristics:** Does not depend on climatological event frequency.
- The expression is identical to *HK* = *POD* - *POFD*, but the Hanssen and Kuipers score can also be interpreted as *(accuracy for events) + (accuracy for non-events) - 1*.
- For rare events *HK* is unduly weighted toward the first term (same as *POD*), so this score may be more useful for more frequent events.
- Can be expressed in a form similar to the *ETS* except the $hits_{random}$ term is unbiased. See Woodcock (1976) for a comparison of *HK* with other scores.

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

$$HK = \frac{hits}{hits + misses} - \frac{false\ alarms}{false\ alarms + correct\ negatives}$$

Hanssen and Kuipers discriminant (HK):Ex 1

Event Observed?

		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

		Yes	No
Forecast	Correct	120	10
	Wrong	20	300

$$HK = \frac{hits}{hits + misses} - \frac{false\ alarms}{false\ alarms + correct\ negatives}$$

$$HK = \frac{120}{120+300} - \frac{10}{10+300} = 82.5\%$$

Hanssen and Kuipers discriminant (HK): Ex 2

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

Forecast		Yes	No
	Correct	120	3000
	Wrong	20	300

$$HK = \frac{hits}{hits + misses} - \frac{false\ alarms}{false\ alarms + correct\ negatives}$$

$$HK = \frac{120}{120+300} - \frac{3000}{3000+300} = -5.2\%$$

-

Heidke skill score (HSS or Cohen's k)

- ***Answers the question:*** *What was the accuracy of the forecast relative to that of random chance?*
- **Range:** minus infinity to 1, 0 indicates no skill.
- **Perfect score:** 1.
- **Characteristics:** Measures the fraction of correct forecasts after eliminating those forecasts which would be correct due purely to random chance.
- This is a form of the generalized skill score, where the *score* in the numerator is the number of correct forecasts, and the reference forecast in this case is random chance.
- In geophysics, random chance is usually not the best forecast to compare to.

Event Observed?

		Yes	No
Forecast	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

$$HSS = \frac{hits - false\ alarms - (expected\ \ correct)_{random}}{Total\ \ Number - (expected\ \ correct)_{random}}$$

$$(expected\ \ correct)_{random} = \frac{1}{N^2}\left[\begin{array}{c}(hits + misses)(hits + \ false\ alarms) + \\ (correct\ \ negs + misses)(correct\ \ negs + \ false\ alarms)\end{array}\right]$$

Heidke skill score (Cohen's k): Example 1

Event Observed?

Forecast		Yes	No
	Correct	Hit	False Alarm
	Wrong	Miss	Correct Negative

Event Observed?

Forecast		Yes	No
	Correct	120	10
	Wrong	20	300

$$expected\ \ correct)_{random} = \frac{1}{N^2}\left[\begin{array}{c}(hits + misses)(hits + \ false\ \ alarms) + \\ (correct\ \ negs + misses)(correct\ \ negs + \ false\ \ alarms)\end{array}\right] = 0.58$$

$$HSS = \frac{hits - \ false\ \ alarms - (expected\ \ correct)_{random}}{Total\ \ Number - (expected\ \ correct)_{random}}$$

$$HSS = \frac{120 - 10 - 0.58}{120 + 20 + 10 + 300 - 0.58} = 0.24$$

Heidke skill score (Cohen's k): Example 2

Forecast	Event Observed? Yes	Event Observed? No
Correct	Hit	False Alarm
Wrong	Miss	Correct Negative

Forecast	Event Observed? Yes	Event Observed? No
Correct	120	3000
Wrong	20	300

$$xpected\ \ correct)_{random} = \frac{1}{N^2}\left[\begin{array}{c}(hits + misses)(hits + \ false\ alarms) + \\ (correct\ \ negs + misses)(correct\ \ negs + \ false\ alarms)\end{array}\right] = 0.126$$

$$HSS = \frac{hits - \ false\ alarms - (expected\ \ correct)_{random}}{Total\ \ Number - (expected\ \ correct)_{random}}$$

$$HSS = \frac{120 - 3000 - 0.58}{120 + 20 + 3000 + 300 - 0.58} = -0.838$$

Skill Scores

- In general, skill scores (*SS*) are defined as

$$SS = \frac{MSE - MSE_{\mathrm{clim}}}{0 - MSE_{\mathrm{clim}}} = 1 - \frac{MSE}{MSE_{\mathrm{clim}}}$$

- A measure of skill (commonly called a skill score, SS) might be the ratio of the unexplained MSE to the climatological MSE:

$$UNEXPLAINED\ \ MSE = MSE - MSE_{\mathrm{clim}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i} - \overline{x}\right)^{2} - \frac{1}{n}\sum_{i=1}^{n}\left(x_{CLIM,\,i} - \overline{x}\right)^{2}$$

- Where *A* is a measure of accuracy (or inaccuracy)

$$SS = \frac{A - A_{\mathrm{ref}}}{A_{perfect} - A_{\mathrm{ref}}} x100\%$$

- This *A* can be any metric. The SS will be the skill in this metric.

Brier Score and Brier Skill Score

- **Range:** 0 to 1. **Perfect score:** 0.
- **Characteristics:** Sensitive to climatological frequency of the event: the more rare an event, the easier it is to get a good *BS* without having any real skill.
  - Negative orientation (smaller score better), which some people find uncomfortable. This can be "fixed" by subtracting *BS* from 1.

$$BS = \frac{1}{N}\sum_i \left(P_i - O_i\right)^2$$

- Where P_i is the forecast probability of an event occurring, and O_i is either one or zero if the event did or did not occur. N is the number of forecasts.
- The Brier Skill Score (BSS) is calculated like other skill scores:

$$BSS = \frac{BS - BS_{reference}}{0 - BS_{reference}}$$