Appendix A

Verification Indices

This appendix highlights a number of verification indices referenced in this document. The indices are also used in international verification via the WMO Integrated Processing and Prediction System (WIPPS) of the World Meteorological Organization (WMO 2023).

A.1 Basic Verification Indices

A.1.1 Mean Error

Mean Error (ME), also called Bias, represents the mean value of deviations between forecasts and verification values, and is defined by

$$\mathrm{ME} \equiv \left(\sum_{i=1}^{n} w_i D_i \right) \Big/ \sum_{i=1}^{n} w_i, \tag{A.1.1a}$$

$$D_i = F_i - A_i, \tag{A.1.1b}$$

$$w_i = \frac{1}{n} \ (\text{or } \cos\phi_i, \text{ and so on}), \tag{A.1.1c}$$

where F_i, A_i, and D_i represent the forecast, the verification value, and the deviation between the forecast and the verification value, respectively. w_i represents the weighting coefficient, n is the number of samples, and ϕ_i is latitude. In general, observational values, initial values or objective analysis values are often used as verification values. When the forecast is fully correct, called *a perfect forecast*, ME is equal to zero.

Calculation of an average over an extensive region such as the Northern Hemisphere requires evaluation with weighting coefficients in consideration of latitude-related differences among areas. By way of example, to evaluate objective analysis in an equirectangular projection, the weighting coefficient "$w_i = 1/n$" is often replaced with the cosine of latitude "$\cos\phi_i$". The other indices in Section A.1 are handled in the same manner.

A.1.2 Root Mean Square Error

Root Mean Square Error (RMSE) is often used to represent forecast accuracy, and is defined by

$$\mathrm{RMSE} \equiv \sqrt{\sum_{i=1}^{n} w_i D_i^2} \Big/ \sqrt{\sum_{i=1}^{n} w_i}, \tag{A.1.2}$$

where D_i represents deviation between the forecast and the verification values in Eq. (A.1.1b), w_i represents the weighting coefficient in Eq. (A.1.1c), and n is the number of samples. Proximity of the RMSE to zero indicates that forecast values are closer to verification values. For a perfect forecast, RMSE is equal to zero. With the components of ME and random error separated, RMSE is expressed as follows:

$$\mathrm{RMSE}^2 = \mathrm{ME}^2 + \sigma_{\mathrm{e}}^2, \tag{A.1.3}$$

where σ_e represents standard deviation (SD) for the deviation D_i, and is given by

$$\sigma_e^2 = \left(\sum_{i=1}^{n} w_i (D_i - \mathrm{ME})^2 \right) \Big/ \sum_{i=1}^{n} w_i. \tag{A.1.4}$$

A.1.3 Anomaly Correlation Coefficient

The anomaly correlation coefficient (ACC) is one of the most widely used measures in the verification of spatial fields (Jolliffe and Stephenson 2003), and represents the correlation between anomalies of forecasts and those of verification values with reference values such as climatological data. ACC is defined as follows:

$$\mathrm{ACC} \equiv \frac{\sum_{i=1}^{n} w_i \left(f_i - \overline{f}\right)(a_i - \overline{a})}{\sqrt{\sum_{i=1}^{n} w_i \left(f_i - \overline{f}\right)^2 \sum_{i=1}^{n} w_i \left(a_i - \overline{a}\right)^2}}, \quad (-1 \leq \mathrm{ACC} \leq 1), \tag{A.1.5}$$

where n is the number of samples, and f_i, $\overline{f}$, a_i and $\overline{a}$ are given by the following equations:

$$f_i = F_i - C_i, \quad \overline{f} = \left(\sum_{i=1}^{n} w_i f_i \right) \Big/ \sum_{i=1}^{n} w_i, \tag{A.1.6a}$$

$$a_i = A_i - C_i, \quad \overline{a} = \left(\sum_{i=1}^{n} w_i a_i \right) \Big/ \sum_{i=1}^{n} w_i, \tag{A.1.6b}$$

where F_i, A_i, and C_i represent the forecast value, the verification value and a reference such as a climatological value, respectively. $\overline{f}$ is the mean of f_i, $\overline{a}$ is the mean of a_i, and w_i represents the weighting coefficient in Eq. (A.1.1c). If the variation pattern of forecast anomalies is perfectly coincident with that of verification anomalies, the ACC will take the maximum value of 1. Conversely, if the variation pattern is completely reversed, it will take the minimum value of -1.

A.1.4 Ensemble Spread

Ensemble Spread is a familiar measure representing the degree of forecast uncertainty in the ensemble forecast. It is the standard deviation of the ensembles as defined by

$$\text{Ensemble Spread} \equiv \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{M} \sum_{m=1}^{M} (F_{m,i} - \bar{F}_i)^2 \right)}, \tag{A.1.7}$$

where M is the number of ensemble members, N is the number of samples, $F_{m,i}$ represents the forecast of the mth member, and $\bar{F}_i$ is the ensemble mean, defined by

$$\bar{F}_i \equiv \frac{1}{M} \sum_{m=1}^{M} F_{m,i}. \tag{A.1.8}$$

Table A.2.1: Schematic contingency table for categorical forecasts of a binary event. The numbers of outcomes in each category are indicated by FO, FX, XO and XX, and N is the total number of events.

	Observed	Not Observed	Total
Forecasted	FO (hits)	FX (false alarms)	$FO + FX$
Not Forecasted	XO (misses)	XX (correct rejections)	$XO + XX$
Total	M	X	N

A.1.5 S1 Score

The S1 Score is often used to measure the degree of error in the depiction of forecast pressure fields, and is defined by

$$\mathrm{S1} \equiv 100 \times \frac{\sum_{i=1}^{n} w_i \left\{ \left|\partial_x D_i\right| + \left|\partial_y D_i\right| \right\}}{\sum_{i=1}^{n} w_i \left[\max\left(\left|\partial_x F_i\right|, \left|\partial_x A_i\right|\right) + \max\left(\left|\partial_y F_i\right|, \left|\partial_y A_i\right|\right) \right]}, \tag{A.1.9}$$

where F_i and A_i represent forecast and verification values, respectively. D_i is the deviation between the forecast and verification values in Eq. (A.1.1b), w_i is the weighting coefficient in Eq. (A.1.1c), and the subscripts x and y denote the differential with respect to x and y, as expressed by

$$\partial_x X = \frac{\partial X}{\partial x}, \quad \partial_y X = \frac{\partial X}{\partial y}. \tag{A.1.10}$$

Lower S1 Scores indicate superior forecasts.

A.2 Verification Indices for Categorical Forecasts

Many meteorological phenomena can be regarded as simple binary events, and related forecasts or warnings are often issued as unqualified statements indicating that such events will or will not occur (Jolliffe and Stephenson 2003). In the verification of forecasts for binary events, outcomes for the targeted phenomenon are distinguished in terms of correspondence between forecasts and observations using a 2×2 contingency table as shown in Table A.2.1.

A.2.1 Contingency Table

In the contingency table, categorical forecasts for a binary event are divided into hits, false alarms, misses and correct rejections (or correct negatives) with numbers expressed as FO, FX, XO and XX, respectively. The total number of events is the sum of numbers for all outcomes, given by $N = FO + FX + XO + XX$. The numbers of observed events and non-observed events are $M = FO + XO$ and $X = FX + XX$, respectively.

A.2.2 Proportion Correct

Proportion Correct (PC) is the ratio of the number of correct events $FO + XX$ to the total number of events N, and is defined by

$$\text{PC} \equiv \frac{FO + XX}{N}, \qquad (0 \leq \text{PC} \leq 1). \tag{A.2.1}$$

Higher PC values indicate higher forecast accuracy.

A.2.3 False Alarm Ratio

The false alarm ratio (FAR) is the ratio of the number of false alarm events FX to the number of forecast events $FO + FX$, and is defined by

$$\text{FAR} \equiv \frac{FX}{FO + FX}, \qquad (0 \leq \text{FAR} \leq 1). \tag{A.2.2}$$

Lower FAR values indicate a lower number of false alarm events. In some cases, the total number N is used as the denominator in Eq. (A.2.2) instead of $FO + FX$.

A.2.4 Undetected Error Rate

The undetected error rate (Ur) is the ratio of the number of miss events XO to the number of observed events M, and is defined by

$$\text{Ur} \equiv \frac{XO}{M}, \qquad (0 \leq \text{Ur} \leq 1). \tag{A.2.3}$$

Lower Ur values indicate a lower number of miss events. In some cases, the total number N is used as the denominator in Eq. (A.2.3) instead of M.

A.2.5 Hit Rate

The hit rate (Hr) is the ratio of the number of hit events FO to the number of observed events M, and is defined by

$$\text{Hr} \equiv \frac{FO}{M}, \qquad (0 \leq \text{Hr} \leq 1). \tag{A.2.4}$$

Higher Hr values indicate a lower number of miss events. The hit rate is used to plot the ROC curve described in Subsection A.3.5.

A.2.6 False Alarm Rate

The false alarm rate (Fr) is the ratio of the number of false alarm events FX to the number of non-observed events X, and is defined by

$$\text{Fr} \equiv \frac{FX}{X}, \qquad (0 \leq \text{Fr} \leq 1). \tag{A.2.5}$$

Lower Fr values indicate a lower number of false alarm events and higher forecast accuracy. The denominator of the false alarm rate differs from that of the false alarm ratio (see Subsection A.2.3). The false alarm rate is also used to plot the ROC curve described in Subsection A.3.5.

A.2.7 Bias Score

The bias score (BI) is the ratio of the number of forecasted events $FO + FX$ to the number of observed events M, and is defined by

$$\mathrm{BI} \equiv \frac{FO + FX}{M}, \qquad (0 \leq \mathrm{BI}). \tag{A.2.6}$$

If the number of forecasted events $FO + FX$ is equal to the number of observed events M, BI will be unity. If BI is larger than unity, the frequency of events is overestimated. Conversely, if BI is smaller than unity, the frequency of events is underestimated.

A.2.8 Climatological Relative Frequency

Climatological relative frequency (P_c) is the probability of occurrence of events estimated from samples, and is defined by

$$\mathrm{P}_c \equiv \frac{M}{N}, \tag{A.2.7}$$

where M is the number of observed events occurring, and N is the total number of events. P_c is derived from the number of observed events, and is independent of forecast accuracy.

A.2.9 Threat Score

The threat score (TS) is an index value focused on hit events. It represents the ratio of the number of hit events FO to the number of events other than correct rejections $FO + FX + XO$, and is defined by

$$\mathrm{TS} \equiv \frac{FO}{FO + FX + XO}, \qquad (0 \leq \mathrm{TS} \leq 1). \tag{A.2.8}$$

If the number of observed events is extremely small (i.e. $N \gg M$, and $XX \gg FO$, FX, or XO), the proportion correct (PC) value will be close to unity due to the the major contribution from the number of non-observed events. The TS is applicable to validation of forecasts accuracy without contribution from correct rejection events. Forecast accuracy rises as the TS value approaches the maximum value of unity. As TS values are often affected by climatological relative frequency, they are not applicable to comparison regarding the accuracy of forecasts validated under different conditions. To address this issue, equitable threat scores are often used for validation.

A.2.10 Equitable Threat Score

The equitable threat score (ETS) is similar to the threat score, but with the removal of contribution from hits by chance in *random forecasts*, and is defined by (Schaefer 1990)

$$\mathrm{ETS} \equiv \frac{FO - S_f}{FO + FX + XO - S_f}, \qquad (-\frac{1}{3} \leq \mathrm{ETS} \leq 1), \tag{A.2.9}$$

and

$$S_f = P_c(FO + FX), \quad P_c = \frac{M}{N}, \tag{A.2.10}$$

where P_c is the climatological relative frequency and S_f is the number of hit events being forecast randomly $FO + FX$ times. Proximity to the maximum value of unity indicates higher forecast accuracy. For random forecasts, the ETS is zero. This metric has a minimum value of $-1/3$ if $FO = XX = 0$ and $FX = XO = N/2$.

A.2.11 Heidke Skill Score

The Heidke skill score (HSS) is used to remove the effects of issues in individual forecasts in consideration of the number of correct events in a random forecast estimated from climatological probability, and is defined by

$$\mathrm{HSS} \equiv \frac{FO + XX - S}{N - S}, \qquad (-1 \leq \mathrm{HSS} \leq 1), \tag{A.2.11}$$

where

$$S = P_c(FO + FX) + Px_c(XO + XX), \tag{A.2.12}$$

and

$$P_c = \frac{M}{N}, \quad Px_c = \frac{X}{N} = 1 - P_c, \tag{A.2.13}$$

where P_c and Px_c are the climatological relative frequencies of observed and non-observed events in random forecasting, respectively. Proximity to the maximum value of unity indicates higher forecast accuracy. The Heidke skill score is zero in random forecasts and unity in perfect forecasts. The index has a minimum value of -1 if $FO = XX = 0$ and $FX = XO = N/2$.

A.2.12 Fractions Skill Score

The fractions skill score (FSS) is an index of how forecast skill varies with spatial scale. In other words, it is a measure to verify forecasted fractional event frequencies. The verification method (Roberts and Lean 2008) is described here.

First, all model and observation data are projected onto the same verification grid. Suitable thresholds (q) are chosen and used to convert the observed (O) and forecast (F) fields into binary fields I_O and I_F. All grid squares exceeding the threshold have a value of 1 and all others a value of 0,

$$I_o = \begin{cases} 1 & (O \geq q) \\ 0 & (O < q) \end{cases} \quad and \quad I_F = \begin{cases} 1 & (F \geq q) \\ 0 & (F < q) \end{cases} \tag{A.2.14}$$

Second, for every grid point in the binary fields obtained from Eq. (A.2.14), computation is performed to determine the fraction of surrounding points within a given square of length n that have a value of 1. These are described by

$$O(n)_{i,j} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_o \left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2} \right], \tag{A.2.15}$$

$$F(n)_{i,j} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_F \left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2} \right]. \tag{A.2.16}$$

Third, the mean square error (MSE) for the observed and forecast fractions from the neighborhood of length n is computed using

$$MSE(n) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O(n)_{i,j} - F(n)_{i,j}]^2. \tag{A.2.17}$$

Here i goes from 1 to N_x, where N_x is the number of columns in the domain, and j goes from 1 to N_y, where N_y is the number of rows. $O(n)_{i,j}$ is the resultant field of observed fractions for the square of length n and $F(n)_{i,j}$ is the resultant field of model forecast fractions. However, the MSE is not in itself very useful because it is highly dependent on the frequency of the event itself. The fractions skill score is defined by

$$FSS(n) = \frac{MSE(n) - MSE(n)_{ref}}{MSE(n)_{perfect} - MSE(n)_{ref}} = 1 - \frac{MSE(n)}{MSE(n)_{ref}} \tag{A.2.18}$$

where $MSE(n)_{perfect}$ is the MSE of a perfect forecast for neighborhood length n and $MSE(n)_{ref}$ is the MSE of the reference.

A.3 Verification Indices for Probability Forecasts

A.3.1 Brier Score

The Brier score (BS) is a basic verification index for probability forecasts, and is defined by

$$\mathrm{BS} \equiv \frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2, \quad (0 \le \mathrm{BS} \le 1), \tag{A.3.1}$$

where p_i is the forecast probability of occurrence of an event ranging from 0 to 1 in probability forecasts, a_i indicates observation with binary values (1 for observed and 0 for not observed), and N is the number of samples. Smaller BS values indicate higher forecast accuracy. In a perfect forecast, the BS has a minimum value of 0.

The Brier score for *climatological forecasts* (BS_c), in which the climatological relative frequency $P_c = M/N$ is always used as the forecast probability p_i ,is defined by

$$\mathrm{BS}_c \equiv P_c(1 - P_c), \tag{A.3.2}$$

Since the Brier score is influenced by the climatological frequency of events in the verification sample, it is not applicable to comparison of accuracy for forecasts with different sets of samples and/or different phenomena. For example, BS_c may differ with differing values of P_c even under the same forecast method (e.g., the climatological approach) because of its dependence on P_c (Stanski and Burrows 1989). To reduce this effect, the Brier skill score is often used for verification with improvement from the climatological forecast (see Subsection A.3.2).

A.3.2 Brier Skill Score

The Brier skill score (BSS) is an index based on the Brier score. It indicates the degree of forecast improvement in reference to climatological forecasts, and is defined by

$$\mathrm{BSS} \equiv \frac{\mathrm{BS}_c - \mathrm{BS}}{\mathrm{BS}_c}, \quad (\mathrm{BSS} \le 1), \tag{A.3.3}$$

where BS is the Brier score and BS_c is the Brier score for the climatological forecast. BSS is unity for a perfect forecast and zero for the climatological forecast. Its value is negative if the forecast error exceeds that of the climatological forecast.

A.3.3 Murphy's Decompositions

To provide deeper insight into the relationship between the Brier score (BS) and the properties of probability forecasts, Murphy (1973) decomposed the score into reliability, resolution and uncertainty terms (Eq. A.3.4a), refered to as Murphy's Decompositions.

Consider the probability of forecasts classified to L intervals. Let the sample number in the lth interval be N_l, and let the number of observed events in N_l be M_l. It follows that $N = \sum_{l=1}^{L} N_l$ and $M = \sum_{l=1}^{L} M_l$. The BS value can therefore be represented with Murphy's decompositions as follows:

$$\mathrm{BS} = \mathrm{Reliability} - \mathrm{Resolution} + \mathrm{Uncertainty}, \tag{A.3.4a}$$

$$\mathrm{Reliability} = \sum_{l=1}^{L}\left(p_l - \frac{M_l}{N_l}\right)^2 \frac{N_l}{N}, \tag{A.3.4b}$$

$$\mathrm{Resolution} = \sum_{l=1}^{L}\left(\frac{M}{N} - \frac{M_l}{N_l}\right)^2 \frac{N_l}{N}, \tag{A.3.4c}$$

$$\mathrm{Uncertainty} = \frac{M}{N}\left(1 - \frac{M}{N}\right), \tag{A.3.4d}$$

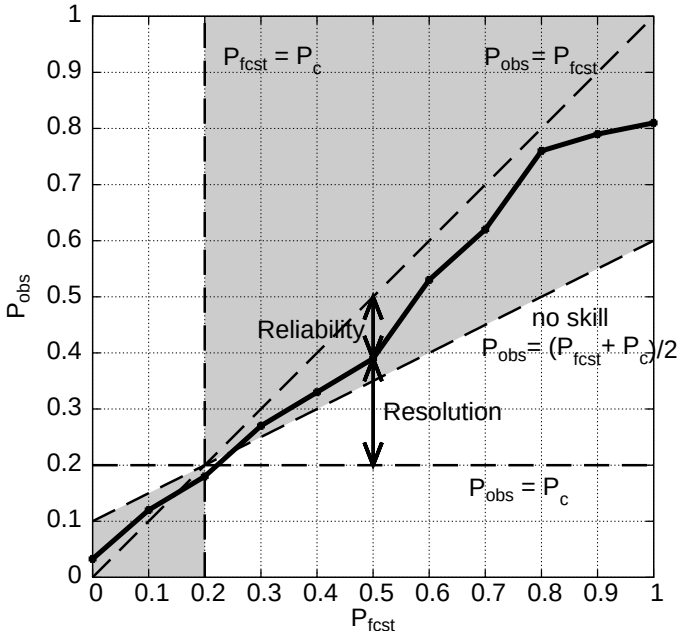

Figure A.3.1: Reliability diagram. The ordinate represents the relative frequencies of observed events P_{obs}, the abscissa is the probability of forecast event occurrence P_{fcst}, and the solid line is the reliability curve. Grey shading indicates positive contribution to the BSS.

where p_l is the representative value in the lth interval of predicted probability. Reliability is the minimum value of zero when p_l is equal to the relative frequency of the observed events M_l/N_l. If the distance between $M/N\,(=P_c)$ and M_l/N_l is longer, resolution will have a large value. Uncertainty depends on observed events regardless of forecast methods. When $P_c = 0.5$, Uncertainty will have the maximum value of 0.25. Uncertainty is equal to the Brier score for climatological forecasts (BS_c). In this regard, the Brier skill score (BSS) can be expressed as

$$\text{BSS} = \frac{\text{Resolution} - \text{Reliability}}{\text{Uncertainty}}. \tag{A.3.5}$$

A.3.4 Reliability Diagram

Probability forecast performance is often evaluated using a reliability diagram, also called an attributes diagram. This is a chart detailing the relative frequencies of observed events P_{obs} as the ordinate and the probability of forecast event occurrence P_{fcst} as abscissa as shown in Figure A.3.1. The plots are generally displayed in the form of a reliability curve.

The properties of the curve can be related to the reliability and resolution terms of Murphy's decompositions. Contribution to reliability (or resolution) for each value of P_{fcst} is associated with the squared distance from a point on the reliability curve to the line $P_{\text{obs}} = P_{\text{fcst}}$ (or $P_{\text{obs}} = P_c$), and is derived from its weighted mean using the number of samples as weights. The contributions are the same for both reliability and resolution on the line $P_{\text{obs}} = (P_{\text{fcst}} + P_c)/2$, called the no-skill line, and contribution to the Brier score is zero on this line. The shading enclosed by the no-skill line, the line $P_{\text{fcst}} = P_c$ and the axes in Figure A.3.1 indicate the area of positive contribution to the BSS, since the contribution to reliability is larger than that to resolution. For further details of reliability diagrams, see Wilks (2006).

In climatological forecasting (see Subsection A.3.1) as a special case, the reliability curve corresponds to a point $(P_{\text{fcst}}, P_{\text{obs}}) = (P_c, P_c)$. Probability forecasts with the following properties will have higher accuracy.

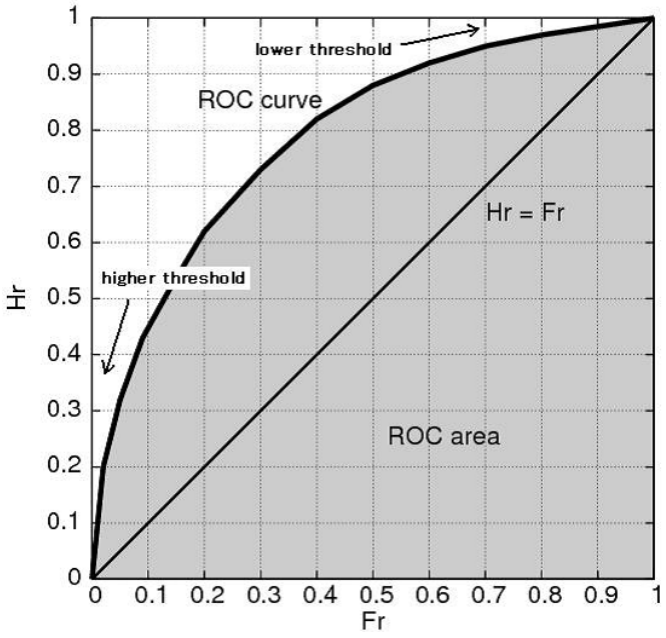

Figure A.3.2: Schematic diagram of an ROC curve. The ordinate is Hr and the abscissa is Fr. Gray shading indicates the ROC area.

- The reliability curve is close to the line $P_{\text{obs}} = P_{\text{fcst}}$ (reliability close to zero).
- Points with a large number of samples on the reliability curve are distributed away from the point of the climatological forecast $(P_{\text{fcst}}, P_{\text{obs}}) = (P_c, P_c)$ (around the lower left or the upper right of the reliability diagram) with higher resolution.

A.3.5 ROC Area Skill Score

If two alternatives in a decision problem, whether the event occurs or not, must be chosen on the basis of a probability forecast for a dichotomous variable, the choice will depend on the probability threshold. A relative operating characteristic (ROC) curve is often used to evaluate such decision problems. This involves the use of a schematic diagram whose ordinate and abscissa represent the hit rate (Hr) and the false alarm rate (Fr), respectively, and is made from contingency tables with variations of threshold values as shown in Figure A.3.2.

The threshold value is lower around the upper right of the diagram and higher around the lower left. Probability forecasting is more accurate when the curve is more convex to the top because the hit rate is higher than the false alarm rate; that is, $Hr > Fr$ around the upper left. The gray shaded area below the ROC curve, called the ROC area (ROCA), will be larger with higher values of information in probability forecasts. For further details of ROC curves, see Wilks (2006).

The ROC area skill score (ROCASS) is a validation index in reference to probability forecasts with no information values (i.e. $Hr = Fr$), and is defined by

$$\text{ROCASS} \equiv 2(\text{ROCA} - 0.5), \quad (-1 \leq \text{ROCASS} \leq 1). \tag{A.3.6}$$

ROCASS is unity for a perfect forecast and zero for a forecast with no information values, such as one with a uniform probability as randomly sampled from the range [1, 0].