Appendix A

Verification Indices

In this appendix, a number of verification indices used in this document are presented for reference. The indices are also used in the international verification through the Global Data-processing and Forecasting System (GDPFS) of the World Meteorological Organization (WMO 2010a, 2012).

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

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

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

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

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

In calculating the average in a wide region, *e.g.* the Northern hemisphere, the average should be evaluated with the weighting coefficients, taking into account the differences of areas due to the latitudes. For example, in order to evaluate objective analysis in equirectangular projection, the weighting coefficient "$w_i = 1/n$" is often replaced with cosine of latitude "$\cos\phi_i$" (see WMO (2012)). The other indices in Section A.1 will be dealt with in the same manner.

A.1.2 Root Mean Square Error

Root Mean Square Error (RMSE) is often used for representing the accuracy of forecasts, and is defined by

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

where D_i represents the deviation between forecast and verifying value in Eq. (A.1.1b), w_i represents the weighting coefficient in Eq. (A.1.1c), and n is the number of samples. If RMSE is closer to zero, it means that the forecasts are closer to the verifying values. For perfect forecast, RMSE is equal to zero. By separating the components of ME and random error, 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

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

A.1.3 Anomaly Correlation Coefficient

Anomaly Correlation Coefficient (ACC) is one of the most widely used measures in the verification of spatial fields (Jolliffe and Stephenson 2003), and is the correlation between anomalies of forecasts and those of verifying values with the reference values, such as climatological values. ACC is defined as follows:

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

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

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

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

where F_i, A_i, and C_i represent forecast, verifying value, and reference value such as climatological value, respectively. Also, $\overline{f}$ 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 the anomalies of forecast is perfectly coincident with that of the anomalies of verifying value, ACC will take the maximum value of 1. In turn, if the variation pattern is completely reversed, ACC takes the minimum value of -1.

A.1.4 Ensemble Spread

Ensemble Spread is a familiar measure which represents the degree of the forecast uncertainty in the ensemble forecast. It is the standard deviation of the ensembles, and is 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

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

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

where F_i and A_i represent forecast and verifying value, respectively. D_i is the deviation between forecast and verifying value in Eq. (A.1.1b), w_i is the weighting coefficient in Eq. (A.1.1c), and the subscript x or y denotes the differential with respect to x or y, as shown in the forms:

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

The lower S1 Score is, the better the forecast is.

A.2 Verification Indices for Categorical Forecasts

Many meteorological phenomena can be regarded as simple binary events, and forecasts or warnings for these events are often issued as unqualified statement that they will or will not take place (Jolliffe and Stephenson 2003). In the verification of the forecasts for binary events, the outcomes for an event on the targeted phenomenon are distinguished in terms of the correspondence between forecasts and observations, using 2 × 2 contingency table as shown in Table A.2.1.

A.2.1 Contingency Table

In the contingency table, the categorical forecasts of a binary event are divided into four possible outcomes, namely, hits, false alarms, misses, and correct rejections (or correct negatives). The numbers of the possible outcomes are indicated with the notations, FO, FX, XO, and XX, respectively. The total number of events is the sum of numbers of all outcomes, given by $N = FO + FX + XO + XX$. The numbers of observed events and not 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

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

The larger PC means the higher accuracy of the forecasts.

A.2.3 False Alarm Ratio

False Alarm Ratio (FAR) is the ratio of the number of false alarm events FX to the number of forecasted events $FO + FX$, and is defined by

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

The smaller FAR is, the less the number of false alarm events is. 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

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

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

The smaller Ur is, the less the number of miss events is. 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

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

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

The larger Hr is, the less the number of miss events is. Hit Rate is used for the plot of ROC curve, described in Subsection A.3.5.

A.2.6 False Alarm Rate

False Alarm Rate (Fr) is the ratio of the number of false alarm events FX to the number of not observed events X, and is defined by

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

The smaller Fr means that the number of false alarm events is less and the accuracy of the forecasts is higher. It is noted that the denominator of False Alarm Rate is different from that of False Alarm Ratio (see Subsection A.2.3). False Alarm Rate is also used for the plotting of the ROC curve, described in Subsection A.3.5.

A.2.7 Bias Score

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 the events estimated from the samples, and is defined by

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

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

A.2.9 Threat Score

Threat Score (TS) is the index focused on the hit event. TS is the ratio of the number of hit events FO to the number of events except for the correct rejections events $FO + FX + XO$, 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, Proportion Correct (PC) will be close to unity because of the the major contribution from the number of not observed events. TS is applicable to validate the accuracy of forecasts without the contribution from the correct rejections events. The accuracy of forecasts is higher as TS approaches to the maximum value of unity. TS is often affected by Climatological Relative Frequency, so that it is not applicable to compare the accuracy of forecasts validated under different conditions. In order to avoid this problem, Equitable Threat Score is often used for the validation.

A.2.10 Equitable Threat Score

Equitable Threat Score (ETS) is similar to the threat score, but removed the contribution from hits by chance in *random forecast*, and is defined by

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

and

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

where P_c is Climatological Relative Frequency, and S_f is the number of hit events in being forecasted randomly at $FO + FX$ times. The closer to the maximum value of unity, the higher the accuracy of forecast is. In the case of random forecast, ETS is zero. ETS has the minimum value of $-1/3$, if $FO = XX = 0$ and $FX = XO = N/2$.

A.2.11 Skill Score

Skill Score, also called Heidke Skill Score, is used to remove the effect of the difficulties in individual forecasts, taking in to account the number of correct events in random forecast estimated from climatological probabilities, and defined by

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

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

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 events and not observed events in random forecast, respectively. The closer to the maximum value of unity, the higher the accuracy of forecast is. Skill score is zero in random forecast and unity in perfect forecast. Skill score has the minimum value of -1, if $FO = XX = 0$ and $FX = XO = N/2$.

A.3 Verification Indices for Probability Forecasts

A.3.1 Brier Score

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

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

where p_i is the forecast probability of occurrence of an event ranging from 0 to 1 in probability forecasts, a_i indicates the observations with binary values, i.e. 1 for observed or 0 for not observed, and N is the number of samples. The smaller BS is, the higher the accuracy of forecasts is. In the perfect forecast, BS has the minimum value of 0 for the deterministic forecast, in which p_i is equal to 0 or 1.

Brier Score for *climatological forecast* (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 the event in the verification sample, it is not applicable to compare the accuracy of the forecast with different sets of samples and/or different phenomena. For example, BS_c can be different with the different value of P_c even if the forecast method is same such as climatological forecast, because of its dependence on P_c. In order to reduce this effect, Brier Skill Score is often used for the verification with the improvement from the climatological forecast (see Subsection A.3.2).

A.3.2 Brier Skill Score

Brier Skill Score (BSS) is an index based on the Brier Score, indicating the degree of forecast improvements in reference to climatological forecast. BSS is defined by

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

where BS is Brier Score, and BS_c is the Brier Score for climatological forecast. BSS is unity for perfect forecast, and zero for the climatological forecast. BSS has a negative value if the forecast error is more than that of climatological forecast.

A.3.3 Murphy's Decompositions

In order to provide a deeper insight on the relation between Brier Score (BS) and the properties of the probability forecasts, Murphy (1973) decomposed the Brier Score into three terms, i.e. reliability, resolution, and uncertainty. This is called Murphy's Decompositions.

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

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

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

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

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

where p_l is the representative value in the lth interval of the predicted probability. Reliability becomes 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 the 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 forecast (BS_c). In this regard, Brier Skill Score (BSS) can be written as

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

A.3.4 Reliability Diagram

The performance for the probability forecasts is often evaluated using Reliability Diagram, also called Attributes Diagram, which is a chart with the relative frequencies of observed events P_{obs} as the ordinate and the probability of the forecasted events to occur P_{fcst} as abscissa, as shown in Figure A.3.1. The plot is generally displayed as a curve, called Reliability Curve.

The properties of Reliability Curve can be related to Reliability and Resolution in Murphy's Decompositions. Contribution to Reliability (or Resolution) for each value of P_{fcst} is associated with the squared distance from a point on 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 no-skill line, and the contribution to Brier Score becomes zero on this line. The gray meshed area surrounded by the no-skill line, the line $P_{\text{fcst}} = P_c$ and the axes in Figure A.3.1 indicates the area with positive contributions to BSS, since the contribution to Reliability is larger than that to Resolution. For further details on Reliability Diagram, please refer to Wilks (2006).

In the climatological forecast (see Subsection A.3.1) as the special case, the Reliability Curve corresponds to a point $(P_{\text{fcst}}, P_{\text{obs}}) = (P_c, P_c)$. The probability forecasts which indicate the following properties will have higher accuracy.

- Reliability Curve is close to the linear line $P_{\text{obs}} = P_{\text{fcst}}$ (Reliability is close to zero),

- Points with the large number of samples on Reliability Curve is distributed apart 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 in Reliability Diagram), with higher Resolution.

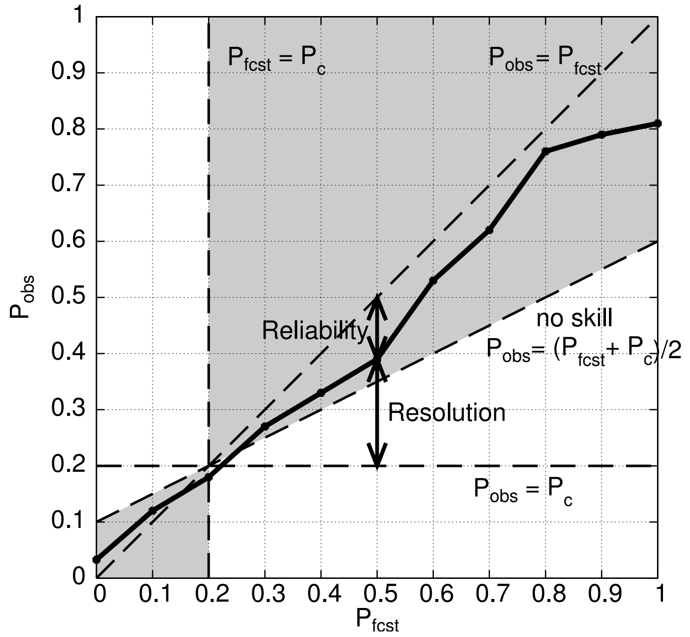

Figure A.3.1: Reliability Diagram. The ordinate is the relative frequencies of observed events P_{obs}, the abscissa is the probability of the forecasted events to occur P_{fcst}, and the solid line is Reliability Curve. The gray meshed area indicates the existence of the positive contributions to BSS.

A.3.5 ROC Area Skill Score

If two alternatives in a decision problem, whether the event occur or not, must be chosen on the basis of a probability forecast for a dichotomous variable, the determination which of the two alternatives will depend on the probability threshold. Relative Operating Characteristic (ROC) curve is often used to evaluate such decision problem. ROC curve is a schematic diagram whose ordinate and abscissa are Hit Rate (Hr) and False Alarm Rate (Fr), respectively, and made from the contingency tables with variations of the threshold values, as shown in Figure A.3.2.

The threshold value is lower around the upper right in the diagram, and higher around the lower left. The probability forecast is more accurate when the curve is more convex to the top because Hit Rate is more than False Alarm Rate, i.e. $Hr > Fr$ around the upper left. Therefore, the area below ROC curve filled in gray, called ROC area (ROCA), will be wider with the higher value of information in the probability forecasts. For further details on ROC curve, please refer to Wilks (2006).

ROC Area Skill Score (ROCASS) is a validation index in reference to the probability forecasts with no value of information, i.e. $Hr = Fr$, and defined by

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

ROCASS is unity for perfect forecast, and zero for the forecast with no value of information, *e.g.* the forecast with a uniform probability which is randomly sampled from the range [1, 0].

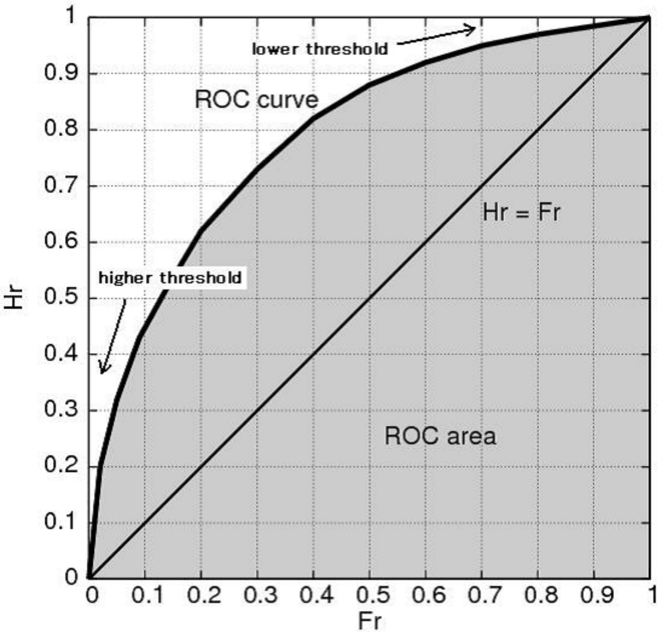

Figure A.3.2: Schematic Diagram of ROC Curve. The ordinate of the diagram is Hr and the abscissa is Fr. The gray area indicates ROC area.