付録B　数値予報課報告・別冊で用いた表記と統計的検証に用いる代表的な指標 *

本報告で使用した表記と統計的検証に用いる代表的な指標などについて以下に説明する。

B.1　数値予報課報告・別冊で用いた表記

B.1.1　GSM, MSM のバージョン名について

気象庁全球モデル (GSM) 及び気象庁メソモデル (MSM) のバージョン名は、GSM, MSM に改良が導入された西暦の下二桁と月を「GSM」や「MSM」の後ろに付けた形式で付けられている（例：GSM1705, MSM1702）。

B.1.2　分解能の表記について

本報告では、全球モデルの分解能について、xx を水平方向の切断波数、yy を鉛直層数として、“TxxLyy”[1] と表記することがある。また、セミラグランジアンモデルで線形格子 (北川 2005) を用いる場合は “TLxxLyy”[2] と表記する。北緯 30 度において、TL959 は約 20 km 格子、TL479 は約 40 km 格子、TL319 は約 55 km 格子、TL159 は約 110 km 格子、TL63 は約 270 km 格子に相当する。

B.1.3　時刻の表記について

本報告では、時刻を表記する際に、通常国内で用いられている日本標準時 (JST: Japan Standard Time) のほかに、協定世界時 (UTC: Coordinated Universal Time) を用いている。数値予報では国際的な観測データの交換やプロダクトの利用などの利便を考慮して、時刻は UTC で表記されることが多い。JST は UTC に対して 9 時間進んでいる。また、単に「時」を用いる場合は、日本標準時を意味する。

B.1.4　予測時間の表記について

数値予報では、統計的な検証や事例検証の結果を示す際に、予報対象時刻のほかに、初期時刻からの経過時間を予報時間 (FT: Forecast Time[3]) として表記している。

本報告では、予報時間を

「予報時間」＝「予報対象時刻」−「初期時刻」

で定義し、例えば、6 時間予報の場合、FT=6 と表記しており、時間の単位 [h] を省略している。

B.1.5　アンサンブル予報の表記について

アンサンブル予報では、複数の予測の集合（アンサンブル）を統計的に処理し、確率予測などの資料を作成する。本報告では、予測の集合の平均を「アンサンブル平均」、個々の予測を「メンバー」と呼ぶ。また、摂動を加えているメンバーを「摂動ラン」、摂動を加えていないメンバーを「コントロールラン」と呼ぶ。全メンバーの数に対する、予測がある閾値を超える（または下回る）メンバーの数の割合を超過確率と呼ぶ。

B.1.6　緯度、経度の表記について

本報告では、緯度、経度について、アルファベットを用いて例えば「北緯 40 度、東経 130 度」を「40°N, 130°E」、「南緯 40 度、西経 130 度」を「40°S, 130°W」などと略記する。

B.2　統計的検証に用いる代表的な指標

B.2.1　平均誤差、二乗平均平方根誤差、誤差の標準偏差、改善率

予測誤差を表す基本的な指標として、平均誤差（ME: Mean Error、バイアスと表記する場合もある）と二乗平均平方根誤差 (RMSE: Root Mean Square Error) がある。これらは次式で定義される。

$$\mathrm{ME} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i) \tag{B.2.1}$$

$$\mathrm{RMSE} \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - a_i)^2} \tag{B.2.2}$$

ここで、N は標本数、x_i は予測値、a_i は実況値である。ME は予測値の実況値からの偏りの平均であり、0 に近いほど実況からのずれが小さいことを示す。RMSE は最小値の 0 に近いほど予測が実況に近いことを示す。

RMSE は ME の寄与とそれ以外を分離して、

$$\mathrm{RMSE}^2 = \mathrm{ME}^2 + \sigma_e^2 \tag{B.2.3}$$

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i - \mathrm{ME})^2 \tag{B.2.4}$$

と表すことができる。σ_e は誤差の標準偏差である。

本報告では、予測に改良を加えた際の評価指標として、RMSE の改善率 (%) を用いる場合がある。RMSE の改善率は次式で定義される。

$$\text{RMSE 改善率} \equiv \frac{\mathrm{RMSE_{cntl}} - \mathrm{RMSE_{test}}}{\mathrm{RMSE_{cntl}}} \times 100 \tag{B.2.5}$$

（RMSE 改善率 ≤ 100）

ここで、$\mathrm{RMSE_{cntl}}$ は基準となる予測の、$\mathrm{RMSE_{test}}$ は改良を加えた予測の RMSE である。

---

* 嶋田 充宏

[1] T は三角形 (Triangular) 波数切断、L は層 (Level) を意味する。

[2] TL の L は線形 (Linear) 格子を意味する。

[3] 英語圏では Forecast Range などと記述されることも多い。

B.2.2　スプレッド

スプレッドは、アンサンブル予報のメンバーの広がりを示す指標であり、次式で定義される。

$$\text{スプレッド} \equiv \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}(x_{mn}-\overline{x_n})^2\right)} \tag{B.2.6}$$

ここで、M はアンサンブル予報のメンバー数、N は標本数、x_{mn} は m 番目のメンバーの予測値、$\overline{x_n}$ は

$$\overline{x_n} \equiv \frac{1}{M}\sum_{m=1}^{M} x_{mn} \tag{B.2.7}$$

で定義されるアンサンブル平均である。

B.2.3　アノマリー相関係数

アノマリー相関係数 (ACC: Anomaly Correlation Coefficient) とは、予測値の基準値からの偏差（アノマリー）と実況値の基準値からの偏差との相関係数であり、次式で定義される。

$$\mathrm{ACC} \equiv \frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(A_i-\overline{A}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2\sum_{i=1}^{N}\left(A_i-\overline{A}\right)^2}} \quad (-1 \le \mathrm{ACC} \le 1) \tag{B.2.8}$$

ただし、

$$X_i = x_i - c_i, \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{B.2.9}$$

$$A_i = a_i - c_i, \qquad \overline{A} = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{B.2.10}$$

である。ここで、N は標本数、x_i は予測値、a_i は実況値、c_i は基準値である。基準値としては気候値を用いる場合が多い。アノマリー相関係数は予測と実況の基準値からの偏差の相関を示し、基準値からの偏差の増減のパターンが完全に一致している場合には最大値の 1 をとり、相関が全くない場合には 0 をとり、逆に完全にパターンが反転している場合には最小値の -1 をとる。なお、アノマリー相関係数や ME, RMSE の解説は、梅津ほか (2013) に詳しい。

B.3　カテゴリー検証で用いる指標

カテゴリー検証では、まず、対象となる現象の有無を予測と実況それぞれについて判定し、その結果により標本を分類する。そして、それぞれのカテゴリーに分類された事例数を基に、予測の特性を検証するという手順を踏む。

B.3.1　分割表

分割表は、カテゴリー検証においてそれぞれのカテゴリーに分類された事例数を示す表（表 B.3.1）である。付録 B.3.2 から B.3.12 に示す各スコアは、表 B.3.1 に示される各区分の事例数を用いて定義される。また、以下では全事例数を N=FO+FX+XO+XX、実況「現象あり」の事例数を M=FO+XO、実況「現象なし」の事例数を X=FX+XX と表す。

表 B.3.1　カテゴリー検証で用いる分割表。FO, FX, XO, XX はそれぞれの事例数を示す。

		実況		計
		あり	なし	
予測	あり	適中 (FO)	空振り (FX)	FO+FX
	なし	見逃し (XO)	適中 (XX)	XO+XX
計		M	X	N

B.3.2　適中率

適中率は、予測が適中した割合であり、次式で定義される。

$$\text{適中率} \equiv \frac{\mathrm{FO}+\mathrm{XX}}{N} \quad (0 \le \text{適中率} \le 1) \tag{B.3.1}$$

最大値の 1 に近いほど予測の精度が高いことを示す。

B.3.3　空振り率

空振り率は、予測「現象あり」の事例数に対する空振り（予測「現象あり」かつ実況「現象なし」）の割合であり、次式で定義される。

$$\text{空振り率} \equiv \frac{\mathrm{FX}}{\mathrm{FO}+\mathrm{FX}} \quad (0 \le \text{空振り率} \le 1) \tag{B.3.2}$$

最小値の 0 に近いほど空振り率が小さいことを示す。本報告では分母を FO+FX としているが、代わりに N として定義する場合もある。

B.3.4　見逃し率

見逃し率は、実況「現象あり」の事例数に対する見逃し（実況「現象あり」かつ予測「現象なし」）の割合であり、次式で定義される。

$$\text{見逃し率} \equiv \frac{\mathrm{XO}}{M} \quad (0 \le \text{見逃し率} \le 1) \tag{B.3.3}$$

最小値の 0 に近いほど見逃し率が小さいことを示す。本報告では分母を M としているが、代わりに N として定義する場合もある。

B.3.5　捕捉率

捕捉率 (H_r: Hit Rate、POD (Probability Of Detection) とも呼ばれる) は、実況「現象あり」のときに予測が適中した割合であり、次式で定義される。

$$H_r \equiv \frac{\mathrm{FO}}{M} \quad (0 \le H_r \le 1) \tag{B.3.4}$$

最大値の 1 に近いほど見逃し率が小さいことを示す。捕捉率は、ROC 曲線（付録 B.4.3）のプロットに用いられる。

B.3.6 体積率

体積率 (V_r: Volume Ratio) は、全事例のうち予測の「現象あり」の事例の割合を示す。

$$V_r \equiv \frac{\mathrm{FO}+\mathrm{FX}}{N} \tag{B.3.5}$$

複数の予測の捕捉率が等しい場合、体積率が小さい予測ほど空振り率が小さい良い予測と言える。

B.3.7 誤検出率

誤検出率 (F_r: False Alarm Rate) は、実況「現象なし」のときに予測が外れた割合である。空振り率 (B.3.2) 式とは分母が異なり、次式で定義される。

$$F_r \equiv \frac{\mathrm{FX}}{X} \quad (0 \le F_r \le 1) \tag{B.3.6}$$

最小値の 0 に近いほど、空振り率が小さく予測の精度が高いことを示す。誤検出率は捕捉率（付録 B.3.5）とともに ROC 曲線（付録 B.4.3）のプロットに用いられる。

B.3.8 バイアススコア

バイアススコア (BI: Bias Score) は、実況「現象あり」の事例数に対する予測「現象あり」の事例数の比であり、次式で定義される。

$$\mathrm{BI} \equiv \frac{\mathrm{FO}+\mathrm{FX}}{M} \quad (0 \le \mathrm{BI}) \tag{B.3.7}$$

予測と実況で「現象あり」の事例数が一致する場合に 1 となる。1 より大きいほど予測の「現象あり」の頻度が過大、1 より小さいほど予測の「現象あり」の頻度が過小であることを示す。

B.3.9 気候学的出現率

現象の気候学的出現率 P_c は、標本から見積もられる「現象あり」の平均的な出現確率であり、次式で定義される。

$$P_c \equiv \frac{M}{N} \quad (0 \le P_c \le 1) \tag{B.3.8}$$

この量は実況のみから決まり、予測の精度にはよらない。予測の精度を評価する際の基準値の設定にしばしば用いられる。

B.3.10 スレットスコア

スレットスコア (TS: Threat Score) は、予測または実況で「現象あり」の場合の予測適中事例数に着目して予測精度を評価する指標であり、次式で定義される。

$$\mathrm{TS} \equiv \frac{\mathrm{FO}}{\mathrm{FO}+\mathrm{FX}+\mathrm{XO}} \quad (0 \le \mathrm{TS} \le 1) \tag{B.3.9}$$

出現頻度の低い現象（$N \gg M$、したがって、XX≫FO, FX, XO となって、予測「現象なし」による寄与だけで適中率が 1 に近い現象）について XX の影響を除いて検証するのに有効である。本スコアは最大値の 1 に近いほど予測の精度が高いことを示す。なお、スレットスコアは現象の気候学的出現率の影響を受けやすく、異なる標本や出現率の異なる現象に対する予測の精度を比較するのには適さない。この問題を緩和するため、次項のエクイタブルスレットスコアなどが考案されている。

B.3.11 エクイタブルスレットスコア

エクイタブルスレットスコア (ETS: Equitable Threat Score) は、前項のスレットスコアが現象の気候学的出現率の影響を受けやすいため、気候学的な確率で「現象あり」が適中した頻度を除いて求めたスレットスコアであり、次式で定義される (Schaefer 1990)。

$$\mathrm{ETS} \equiv \frac{\mathrm{FO}-S_f}{\mathrm{FO}+\mathrm{FX}+\mathrm{XO}-S_f} \quad \left(-\frac{1}{3} \le \mathrm{ETS} \le 1\right) \tag{B.3.10}$$

ただし、

$$S_f = P_c(\mathrm{FO}+\mathrm{FX}) \tag{B.3.11}$$

である。ここで、S_f は「現象あり」をランダムに FO+FX 回予測した場合（ランダム予測）の「現象あり」の適中事例数である。本スコアは、最大値の 1 に近いほど予測の精度が高いことを示す。また、ランダム予測で 0 となり、FO=XX=0, FX=XO=$N/2$ の場合に最小値 $-1/3$ をとる。

B.3.12 スキルスコア

スキルスコア (Skill Score) は気候学的確率などによる予測の難易を取り除いて、予測の技術力を評価する指数であり、一般に次式のように定義される。

$$\text{スキルスコア} \equiv \frac{S_{\mathrm{fcst}}-S_{\mathrm{ref}}}{S_{\mathrm{pfct}}-S_{\mathrm{ref}}} \tag{B.3.12}$$

ここで、$S_{\mathrm{fcst}}, S_{\mathrm{pfct}}, S_{\mathrm{ref}}$ は、評価対象の予測・完全予測・比較の基準となる予測（気候学的確率など）の各スコア（適中率）である。本スコアは、最大値の 1 に近いほど予測の精度が高いことを示し、比較の基準となる予測よりも精度が劣る場合、負の値となる。

代表的なスキルスコアは Heidke のスキルスコア (HSS: Heidke Skill Score) で、気候学的な確率で「現象あり」および「現象なし」が適中した頻度を除いて求める適中率であり、次式で定義される。

$$\mathrm{HSS} \equiv \frac{\mathrm{FO}+\mathrm{XX}-S}{N-S} \quad (-1 \le \mathrm{HSS} \le 1) \tag{B.3.13}$$

ただし、

$$S = P_c(\mathrm{FO}+\mathrm{FX}) + P_x(\mathrm{XO}+\mathrm{XX}),$$

$$P_x = \frac{X}{N} \tag{B.3.14}$$

である。ここで、P_x は「現象なし」の気候学的出現率、S は「現象あり」を FO+FX 回（すなわち、「現象なし」を残りの XO+XX 回）ランダムに予測した場合（ランダム予測）の適中事例数である。HSS は、最大値の 1 に近づくほど精度が高く、ランダム予測で 0 となり、FO=XX=0, FX=XO=$N/2$ の場合に最小値 -1 をとる。前項のエクイタブルスレットスコアもスキルスコアの一つで、Gilbert Skill Score とも呼ばれている。

B.4 確率予測に関する指標など

B.4.1 ブライアスコア

ブライアスコア (BS: Brier Score) は、確率予測の統計検証の基本的指標である。ある現象の出現確率を対象とする予測について、次式で定義される。

$$\mathrm{BS} \equiv \frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2 \quad (0 \le \mathrm{BS} \le 1) \tag{B.4.1}$$

ここで、p_i は確率予測値（0 から 1）、a_i は実況値（現象ありで 1、なしで 0）、N は標本数である。BS は完全に適中する決定論的な（p_i=0 または 1 の）予測（完全予測と呼ばれる）で最小値の 0 をとり、0 に近いほど予測の精度が高いことを示す。また、現象の気候学的出現率 P_c（(B.3.8) 式）を常に確率予測値とする予測（気候値予測と呼ばれる）のブライアスコア $\mathrm{BS_c}$ は

$$\mathrm{BS_c} \equiv P_c(1 - P_c) \tag{B.4.2}$$

となる。ブライアスコアは、現象の気候学的出現率の影響を受けるため、異なる標本や出現率の異なる現象に対する予測の精度を比較するのには適さない。例えば上の $\mathrm{BS_c}$ は P_c 依存性を持ち、同じ予測手法（ここでは気候値予測）に対しても P_c の値に応じて異なる値をとる (Stanski et al. 1989)。この問題を緩和するため、次項のブライアスキルスコアが考案されている。

B.4.2 ブライアスキルスコア

ブライアスキルスコア (BSS: Brier Skill Score) は、ブライアスコアに基づくスキルスコアであり、通常気候値予測を基準とした予測の改善の度合いを示す。本スコアは、ブライアスコア BS、気候値予測によるブライアスコア $\mathrm{BS_c}$ を用いて

$$\mathrm{BSS} \equiv \frac{\mathrm{BS_c} - \mathrm{BS}}{\mathrm{BS_c}} \quad (\mathrm{BSS} \le 1) \tag{B.4.3}$$

で定義され、完全予測で 1、気候値予測で 0、気候値予測より誤差が大きいと負となる。

B.4.3 ROC 曲線、ROC 面積、ROC 面積スキルスコア

現象の予測出現確率にある閾値を設定し、これを予測の「現象あり」「現象なし」を判定する基準とすることが可能である。様々な閾値それぞれについて作成した分割表を基に、閾値が変化したときの F_r–H_r 平面上の軌跡をプロットしたものが ROC 曲線（ROC curve: Relative Operating Characteristic curve、相対作用特性曲線）である（図 B.4.1 参照、高野 2002 などに詳しい）。平面内の左上方の領域では $H_r > F_r$ であり、平面の左上側に膨らんだ ROC 曲線特性を持つ確率予測ほど精度が高いものと見なせる。したがって、ROC 曲線から下の領域（図 B.4.1 灰色の領域）の面積（ROCA: ROC Area、ROC 面積）は、情報価値の高い確率予測ほど大きくなる。ROC 面積スキルスコア (ROCASS: ROC Area Skill Score) は、情報価値のない予測 ($H_r = F_r$) を基準として ROC 面積を評価するものであり、次式で定義される。

$$\mathrm{ROCASS} \equiv 2(\mathrm{ROCA} - 0.5) \quad (-1 \le \mathrm{ROCASS} \le 1) \tag{B.4.4}$$

本スコアは、完全予測で最大値の 1 をとる。また、情報価値のない予測（例えば、区間 $[0, 1]$ から一様ランダムに抽出した値を確率予測値とする予測など）では 0 となる。

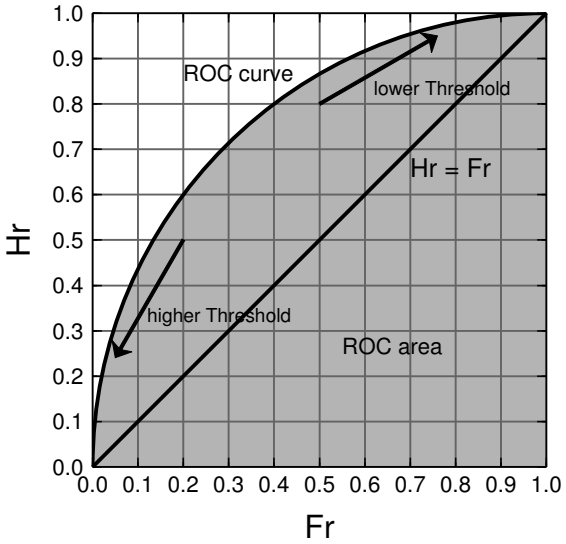

図 B.4.1 ROC 曲線の模式図。横軸は F_r、縦軸は H_r である。灰色の領域の面積が ROC 面積である。

参考文献

北川裕人, 2005: 全球・領域・台風モデル. 平成 17 年度数値予報研修テキスト, 気象庁予報部, 38–43.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in

meteorology. *Research Rep.*, **89-5**, Forecast Research Division, Atmospheric Environment Service, Environment Canada, 114 pp.
高野清治, 2002: アンサンブル予報の利用技術. 気象研究ノート, **201**, 73–103.
梅津浩典, 室井ちあし, 原旅人, 2013: 検証指標. 数値予報課報告・別冊第 59 号, 気象庁予報部, 6–15.