

降水概率预报评分方法分析

周 兵¹, 郭幼君², 何金海¹

(1. 南京气象学院大气科学系, 南京 210044; 2. 北京市气象台, 北京 100089)

摘要: 分析了国内外概率预报评分的几种主要方法及技巧评分过程, 并在预报概率与观测频率(FPOF)分布曲线的基础上提出一种适合我国降水概率预报评估业务的可信度指数方法。论文明确指出1995年北京0.1mm降水概率预报在12~24h和24~36h时段内基本可信, 且有一定的预报技巧; 而10.0mm降水可信度较低, 其24~36h的预报在北京西北部山区布莱尔技巧评分 $B_s < 0$ 。分析表明降水气候概率低的单站技巧评分会出现偏差。

关键词: 预报概率; 观测频率; 评分; 气候频率; 可信度指数; FPOF

中图分类号: P459.9 **文献标识码:** A

早在本世纪20年代就有人提出天气概率预报的初步设想, 1950年布莱尔^[1](Brier)正式提出评估概率的评分方法。60年代美国国家气象局最先把概率预报用于日常气象业务, 并于1965年开始进行降水概率预报试验。70年代以后, 日本和欧洲一些国家和地区相继开展概率预报工作^[2]。1995年夏季我国在北京等地也正式推行降水概率预报业务。期间墨菲(Murphy)等一直致力于概率预报的实施和预报评分及经济效益的评估分析, 并提出相应的方法和措施^[3,4]。1995年丁金才^[5]对天气预报评分方法进行了评述和比较。北京降水概率预报的实践^[6,7]表明降水概率预报较传统定性预报有显著的不同: 它能充分反映降水的不确定性; 引导用户定量地使用降水预报结果; 能发挥天气预报潜在的经济效益。

本文首先结合实际例子对国内外有关概率预报的方法进行系统分析, 在FPOF可信度分布的基础上, 提出一种新的可信度指数, 直接用于北京降水概率预报的评估。

1 概率预报主要评分方法

1.1 布莱尔评分

布莱尔评分是一种均方概率差, 一般地可用下式表示,

$$B = \frac{1}{n} \sum_{i=1}^n (f_i - o_i)^2. \quad (1)$$

其中 n 为两分类事件预报次数; f_i 为事件发生的预报概率; o_i 为事件实际发生情况, 即发生为1, 否则为0。 B 评分是一种逆向评分方法, 其值越小, 预报质量越高, $B=0$, 表示预报完全正确; 反之, 其值越大, 预报质量越低, $B=1$, 表示预报完全不正确。

对于一套假设的10次预报 $f_i = (0.7, 0.9, 0.8, 0.4, 0.2, 0.0, 0.0, 0.0, 0.0, 0.1)$, 与此对应的实际情况是 $o_i = (0, 1, 1, 1, 0, 0, 0, 0, 0, 0)$, 由(1)式可得 $B = 0.095$ 。如果某种事件出现的气候频率为 f_c , 同样可得到气候布莱尔评分值 $B_c = f_c - f_c^2$ 。但布莱尔评分方法并不适用于多分类(multiple-category)事件, 因此需要考虑分类数 k , 此种方法简称为 BM 评分, 用下式表示,

$$BM = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^k (f_{ij} - o_{ij})^2. \quad (2)$$

如果某地降水现象可分成无雨、小雨、中雨和大雨4类, 3个不同预报结果分别为 $A: (0.1, 0.3, 0.5, 0.1)$; $B: (0.5, 0.3, 0.1, 0.1)$ 和 $C: (0.0, 0.1, 0.9, 0.0)$, 而实际情况为大雨, $o_i: (0.0, 0.0, 0.0, 1.0)$, A 、 B 和 C 的 BM 评分结果分别为0.58、0.58和0.91, 所有得分均比较糟。由于预报 C 的高概率值出现在临近类别中雨内, 其结果应该比较接近实际, 其次为预报 A 和预报 B , 但实际评分不能反映出来, 这主要是因 BM 评分对类别间距离不敏感而造成的。

布莱尔评分强烈地依赖于检验样本的气候概率, 对于气候概率为30%的某种预报对象, 最好策略是任何时候的预报概率均为30%, 因此 B 的期望值 $E(B) = 0.21$, 而对诸如雷暴等气候为5%的小概率事件 $E(B) = 0.047$, 这意味着在没有任何预报技巧的情况下, 小概率事件能得到较高的布莱尔评分值, 因此对不同气候概率的事件进行布莱尔评分比较是有风险的。另外, 布莱尔评分方法仍受定性预报的影响, 对于一套完全正确的概率预报 $B \neq 0.0$, 只有在极端情况 f_i 非0即1时才一致。

1.2 墨菲评分

墨菲评分方法(M)是将布莱尔评分划分成意义比较明确的3个部分, 从而有利于对评分值的理解。假定预报概率为0%~100%间以10%的间隔离散分布, 因此预报概率值可划分成 $T = 11$ 个级别, n 次预报的布莱尔评分写成下式,

$$B = \frac{1}{n} \sum_{k=1}^T \sum_{i=1}^{M_K} (f_k - o_{ki})^2. \quad (3)$$

(3)式可根据预报级别为 K 时事件发生的平均频率 $\bar{o}_K$ 和整个事件发生的平均频率 $\bar{o}$ 改写为

$$M = \frac{1}{n} \sum_{k=1}^T \sum_{i=1}^{M_K} [(f_k - \bar{o}_K)^2 - (\bar{o}_K - \bar{o})^2 + (\bar{o} - o_{ki})^2]. \quad (4)$$

上式右端第一项表示可靠性, 第二项表示分辨率, 第三项表示不确定性。对一套已知的预报, 不确定因子是固定的, 可靠性和分辨率项可通过改变概率预报进行调整, 可靠性评分的贡献要大于分辨率项的贡献。大多数情况下, M 一般在0.10~0.25间, 若 $M > 0.30$, 表示预报质量低劣。对小概率事件, 评分效果要好些, 通常低于0.1。

1.3 秩概率评分

秩概率评分(RPS)是由墨菲提出, Epstein(1969)^[8]发展起来的, 并又经过墨菲^[9, 10]进行修改完善。对于有序的多分类概率预报, 需要克服布莱尔评分对类别距离不敏感的缺点, 引入秩概率评分,

$$RPS(p, d) = 1 - \frac{1}{k-1} \left[\sum_{i=1}^k \left(\sum_{n=1}^i p_n - \sum_{n=1}^i d_n \right)^2 \right]. \quad (5)$$

其中 $p: (p_1, p_2, \dots, p_k)$ 和 $d: (d_1, d_2, \dots, d_k)$ 分别表示不同类别 k 的预报概率和实际观测值; d_i 是唯一确定的, 非1即0, 且只能有一个为1。考虑一气温概率预报, 假设将其分为4个类别, $T \leq 30.0$; $30.0 < T \leq 34.0$; $34.0 < T \leq 38.0$ 和 $T > 38.0$ 。两个预报结果分别为 $(0.1, 0.3, 0.5, 0.1)$ 和 $(0.5, 0.3, 0.1, 0.1)$, 根据实际气温出现的类别不同, 可得到不同的 RPS 值

(表1)。类别4与上述 BM 评分结果相比较, RPS 对类别间距离是不敏感的。

表1 不同类别 RPS 评分结果

观测结果		预报概率	
类别	实际值	(0.1, 0.3, 0.5, 0.1)	(0.5, 0.3, 0.1, 0.1)
1	(1, 0, 0, 0)	0.61	0.90
2	(0, 1, 0, 0)	0.87	0.90
3	(0, 0, 1, 0)	0.94	0.70
4	(0, 0, 0, 1)	0.67	0.43

RPS 的主要特征是: 正向性、总体性和敏感性。其取值范围为 $[0, 1]$, 数值越大, 预报越精确。 RPS 能区分布莱尔评分不能区分的类别差异, 对不同类别的误差有指示性。

1.4 技巧评分

任何一种评分结果均可根据它与气候预报或完全预报的比较来检验其技巧性, 即进行技巧评分,

$$B_s = (1 - \frac{B}{B_c}) \times 100 \%。 \tag{6}$$

秩技巧评分 RPS_s 为

$$RPS_s = \frac{RPS - RPS_c}{1 - RPS_c} \times 100 \%。 \tag{7}$$

RPS_c 可以是气候预报或完全预报, 而 RPS_s 的变化范围为 $(-\infty, 1]$ 。上述技巧评分的几个基本特点是: 非严格性、高度非线性 and 相对指示性。

1995年夏季北京降水概率预报的布莱尔技巧评分结果表明: 12~24h 北京城区、西南部山区和东北部山区0.1mm 降水 B_s 值在0.3~0.4之间, 而东南部平原和西北部山区的 B_s 在0.2左右; 另一个时段24~36h 的技巧评分效果稍差, 平均值为0.209。10.0mm 以上降水也有相似的趋势, 北京城区、西南部山区和东北部山区 B_s 值最大, 12~24h 预报明显优于24~36h, 西北部山区预报效果最差, 24~36h $B_s < 0$ 。国外概率预报评分检验已积累几十年的资料, 作为比较, 分析加拿大 Saskatoon 站1982~1989年逐月36h POP B_s 分布曲线 (图1) 发现, 难于辨认技巧评分变化趋势。1985年1月, $B_s = 0.45$, 优于长期气候概率预报, 而2月 $B_s = -0.65$, 比长期气候预报要糟得多。这种 B_s 月振荡客观地存在, 从而影响概率预报的质量。但经半年平滑后其总的变化趋势分析发现80年代初期平均 $B_s < 0$, 而80年代中期以后以 $B_s > 0$ 为主。同时需要说明的是对诸如 Saskatoon 这样降水气候频率低的单站逐月计算 B_s 值会出现偏差, 建议适当

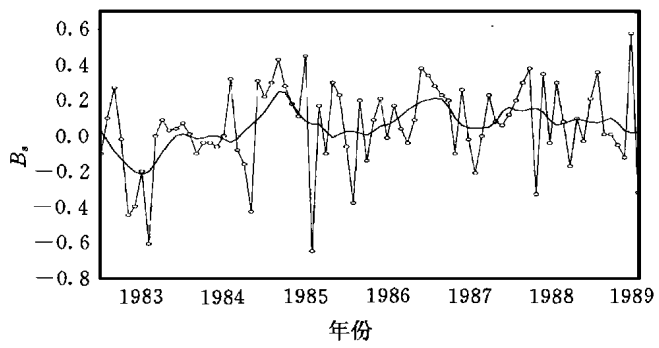


图1 Saskatoon 站逐月降水概率预报
布莱尔技巧评分及滑动平均^[4]

Fig. 1 Monthly Briers kill scores and its running mean for POP forecasts at Saskatoon (after Henry et al.)

延长评分时段。

2 可信度指数及其应用

2.1 可信度分布图

可信度分布图是指预报概率与实际观测频率(FPOF)曲线,它比较直观地反映概率预报的可信程度。图中对角线为预报概率与实际观测频率相等,说明概率预报是完全正确的,对角线下方表示预报有过度行为(overforecast),对角线上方表示预报不足(underforecast)。为了说明问题,假设气候概率 $f_c=0.3$,对一组理想的FPOF分布(图2a~2d)进行分析。

图2a显示在大于气候概率的高概率区,预报有过度行为,而低概率区则相反,预报有不足行为。图2b表明FPOP曲线在预报概率为极端情况($P=0$ 或 1)和气候频率附近可信度得到提高,但依然存在高概率区的过度预报和低概率区的不足预报。图2c为一保守预报系统,在 $P_i \leq 0.6$ 时,预报结果与完全正确的预报(对角线)相一致,但在高概率区,由于不敢进行预报或预报能力的局限,导致 0.7 以上预报概率值使用的观测频率为 0 。这种情况在实践中发生在对小概率事件的预报。图2d是一不稳定系统,由于预报样本数偏少,使得预报结果出现波动,不能有效地反映变化趋势,因此需要适当延长样本长度。

图3为北京1995年夏秋(6~11月)主观POP预报概率-观测频率分布曲线。 0.1mm 降水(图3a)在 $P_i \geq 0.7$ 的高概率区基本正确,但在中间概率段存在严重的过度预报,12~24h和24~36h的趋势一样,在 $P_i=0.5$ 处,观测频率低于气候概率($f_c=0.24$)。 10.0mm 降水(图3b)的FPOF分布具有两个特点:过度预报性和不稳定性。造成这种现象的主要原因是统计样本偏少,北京地区12小时中雨以上降水平均为10次左右。另外也与模式产品对中雨预报释用质量的

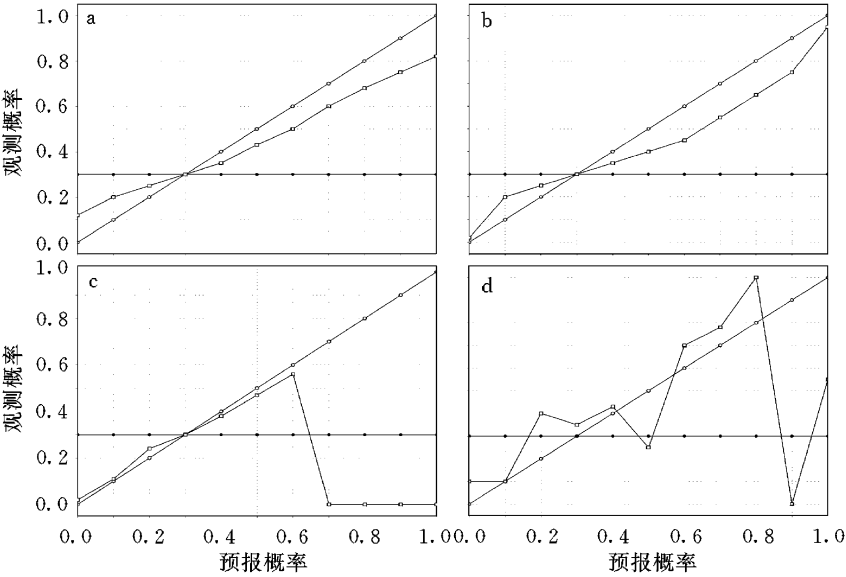


图2 几种假设的预报概率-观测频率分布曲线

●. 气候概率; ○. 完全预报; □. 假想的预报

Fig. 2 Several ideal forecasts probability-observed frequency distribution

●. climatology; ○. perfect prediction; □. ideal reliability curve

下降有关, 而中等预报概率的过度使用主要是缺乏经验造成的。

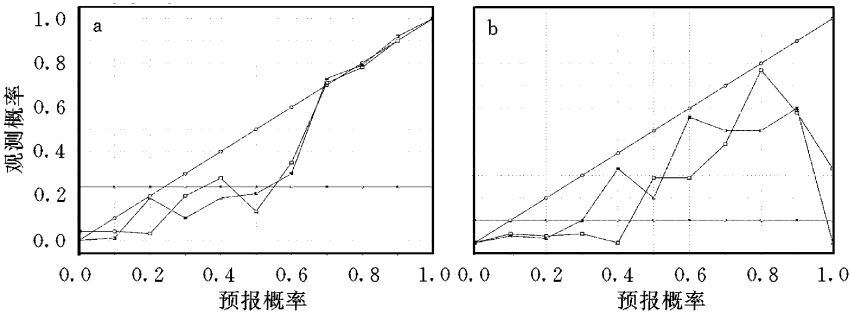


图3 北京1995年6~11月主观 POP 预报概率-观测频率分布曲线
a. 0.1 mm; b. 10.0 mm

●. 气候频率; ○. 完全预报; □. 12~24h 预报; ■. 24~36h 预报

Fig. 3 Subjective POP FPOF distribution curves

a. 0.1 mm; b. 10.0 mm

●. climatology; ○. perfect prediction; □. 12~24h forecasts; ■. 24~36h forecasts

2.2 可信度指数的提出与应用

概率预报验证评分方法的研究已有近50年的历史, 但自1995年我国正式推行降水概率预报业务至今尚未确立能用于业务考核的概率预报评分方法。一般地任何一种评分方法的构造都是主观的、多元的, 但评分结果应该是客观的、全面的和有意义的。为此, 评分方法因遵循以下若个准则^[5]: 1) 构造一个简单易懂的值来表示评分结果; 2) 构造评分值时应包括列联表中的所有元素; 3) 该值与事件本身概率是独立的, 使评分结果有可比性; 4) 避免导向预报偏于某种错误倾向; 5) 反映预报能力的变化有一定的灵敏度; 6) 能用统计假设值对预报技巧的显著性进行检验; 7) 对预报的改进方向有一定的指示意义。在可信度分布曲线的基础上, 提出可信度定量值可信度指数 R。为了更好地说明 R 的计算过程, 现用图4表示。

在 FPOF 图上, 对于预报概率集 $P = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, 有相应的观测频率集 $O = \{0.1, 0.2, 0.1, 0.2, 0.3, 0.2, 0.5, 0.8, 0.7, 0.8, 0.8\}$, 即实曲线 $AB \dots K$ 。设 $\Delta_i = P_i - O_i$, 对 FPOF 分布曲线可能出现的3种情况加以讨论。1) 若 $\Delta_i > 0$, 表示预报过度; 2) 若 $\Delta_i < 0$, 表示预报不足; 3) 若 $\Delta_i = 0$, 预报完全正确。预报概率 P_i 的可信度为 $R_i = O_i / P_i, i = 2, 11$ 。对于 $P_1 = 0.0, R_1 = 1 - O_1$ 。可信度集 R

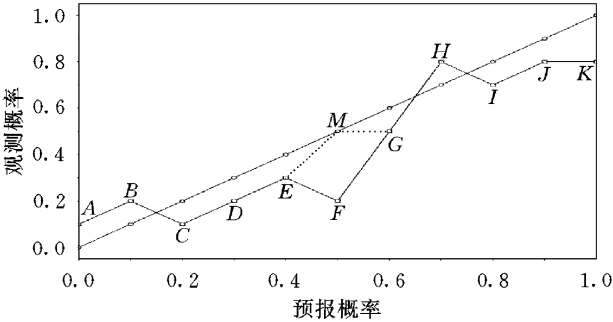


图4 可信度指数分析示意图

Fig. 4 Schematic diagram of reliability index analysis

$= \{0.9, 0.0, 0.5, 0.67, 0.75, 0.40, 0.83, 0.86, 0.87, 0.89, 0.80\}$, 由此可计算总体可信度指数 R 。S 表示曲线 $AB \dots K$ 与对角线预报所围总面积, $S = S_1 + S_2$, S_1 表示对角线上方 FPOF 分布曲线所围面积, S_2 表示对角线下方 FPOF 分布曲线所围面积。 $R = 1 - 2S_2$, 由此可得到上述观测频率集 O 的可信度指数为 $R = 0.78$ 。如果预报概率在 $P_6 = 0.5$ 处即 M 点能得到改进(见图

虚线部分), 当 $O_6 = 0.5$ 时, $\Delta_6 = 0$, 则 $R = 0.84$, 可信度指数提高了 7.6%, 由此可见, 预报概率值应用的提高可反映预报能力的变化, 同时预报的改进方向有明确的指示意义。

对 1995 年北京降水概率预报结果(图 3) 进行可信度指数计算, 可以得到 0.1 mm 降水 12 ~ 24 h 和 24 ~ 36 h 的 R 分别为 0.786 和 0.776; 10.0 mm 降水 12 ~ 24 h 和 24 ~ 36 h 的 R 分别为 0.529 和 0.568, 基本反映了目前北京地区主观降水概率预报水平。结合实际业务可把 R 分成 5 个等级: $R < 0.50$, 预报不可信; $0.5 \sim 0.69$, 可信度较低; $0.70 \sim 0.79$, 基本可信; $0.80 \sim 0.89$, 可信度较好; ≥ 0.90 , 完全可信。

3 结论和问题

(1) B 评分作为概率预报的主要评分方法已广泛地应用于国内外业务预报质量的考核中, 其强烈地依赖于检验样本的气候概率, 具有逆向性, 且对于不同类别 B 值不能进行有效的比较, 对多分类预报类别间距离不敏感。同时, B 评分仍属定性预报范畴, 对一套完全正确的概率预报只有在极端情况下 $B = 0$ 。

(2) M 评分可明确区分不确定性、可靠性和分辨率项对 B 评分的贡献, 在实际业务中尚未普及, 且同样存在 B 评分的弊端。秩概率评分具有正向性、总体性和敏感性等特点, 能有效地反映不同有序类别之间预报概率对 RPS 的影响, 适用于多分类概率预报的评估。

(3) 可信度分布图直观、定性地反映了概率预报的可信程度, 在此基础上提出的可信度指数定量地表征这种程度, 此方法遵循评分的若干准则, 对我国现阶段降水概率预报的业务评分会有重要作用。1995 年北京 0.1 mm 降水预报基本可信, 而 10.0 mm 降水的可信度较低。

(4) 北京夏季 12 ~ 24 h POP 预报技巧优于 24 ~ 36 h, 80 年代国外的逐月 B_s 也存在明显的振荡, 但总体水平有提高。技巧评分具有非严格性、高度非线性性和相对指示性。

致谢: 第一作者对 1995 年降水概率预报业务实施中课题组雷振发、陆晨等和北京市气象台的支持表示衷心感谢。

参 考 文 献

- [1] BRIER G W. Verification of forecasts expressed in terms of probability. Mon Wea Rev, 1950, 78(1): 1 ~ 3
- [2] 周 兵. 国内外降水概率预报研究及业务应用概况[J]. 北京气象, 1995, (3): 8 ~ 12
- [3] MURPHY A H. A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation[J]. J Appl Meteor, 1966, 5(5): 534 ~ 537
- [4] HENRY R S, WILSON L J, BURROWS W R. Survey of common verification methods in meteorology[R]. Research report No. 89-5, Canada, 1991. 43 ~ 64
- [5] 丁金才. 天气预报评分方法评述[J]. 南京气象学院学报, 1995, 18(1): 143 ~ 150
- [6] 周 兵, 陆 晨, 周小平. 北京地区夏季降水概率预报业务应用研究[J]. 气象, 1996, 22(1): 3 ~ 6
- [7] 周 兵, 雷振发. 北京降水概率预报业务试验[J]. 大气科学, 1998, 22(3): 336 ~ 342
- [8] EPSTEIN E S. A scoring system for probability forecasts of ranked categories[J]. J Appl Meteor, 1969, 8(4): 985 ~ 987
- [9] MURPHY A H. On the "ranked probability score" [J]. J Appl Meteor, 1969, 8(4): 988 ~ 989
- [10] MURPHY A H. A note on the ranked probability score. [J] J Appl Meteor, 1971, 10(1): 155 ~ 157

ANALYSIS OF POP FORECAST'S VERIFICATION TECHNIQUES

ZHOU Bing¹, GUO You-jun², HE Jin-hai¹

(1. Department of Atmospheric Sciences, NIM, Nanjing 210044; 2. Beijing Meteorological Observatory, Beijing 100089)

Abstract: This paper investigates the several verification techniques and their skills cores about probability of precipitation (POP), and presents a new reliability index which will be adapted to Chinese POP operation in accordance with the forecast probability and observed frequency's distribution curve. Verification results show that the POP forecasts in Beijing in 1995 are reliable and skillful for light rain, but lesser reliable for 24~36 h forecasts of moderate rain. However, the analysis suggests that the skill scores would be biased in the place wherein the climatic frequency of precipitation is low.

Keywords: forecast probability; observed frequency; verification techniques; climatology frequency; reliability index; FPOF

简 讯

《南京气象学院学报》在教育部和江苏省组织的期刊评比中获奖

接教育部1999年7月1日通知,我院学报在“1999年全国优秀高校自然科学学报及教育部优秀科技期刊评比”中荣获二等奖。此前,学报还在“第三届江苏省优秀期刊”评比中被评为江苏省优秀期刊,也是学报连续第三次获此殊荣。

近年来,《南京气象学院学报》在院领导和主管部门领导的关心支持下,在编委会的直接领导下,在学报编辑、出版工作的各方面进行了改革,严格执行国家科技期刊管理条例和有关国家标准,努力提高学报学术质量、编排质量及印刷质量,取得了一定成效,并得到了各方面的认可。学报在支持学院科研活动、扩大学术交流、推广新技术和新方法等方面起到了一定作用,为学院的发展作出了贡献。

学报编辑部