国家气象中心集合数值预报检验评价*

皇甫雪官
(国家气象中心 ,北京 100081)

提　　要

对国家气象中心的奇异向量法初值扰动的 32 个成员的集合预报 ,利用 2000 年 5 月到 10 月的实况资料进行了 Talagrand 概率分布 、离散度 、Brier 评分(BS) 、Brier 技巧评分(BSS) 、命中率以及空报率的统计检验 。并且对奇异向量法和时间滞后法的两种集合预报结果也进行了比较分析 。

关键词 : 集合预报　检验　比较分析

引　言

由于观测误差和资料处理 、同化分析中引入的误差 ,我们所得到的作为数值预报模式初值的初始场总是含有不确定性 。而数值预报模式的方程组是高度非线性系统 ,对初值的误差具有较强的敏感性 ,预报时间越长误差就成倍增长 。因此 ,某一初值状态下该模式的数值预报的解可能是错的 。为了找出所有可能的解 ,首先就要估算出初值中误差分布的可能范围 ,根据这一范围 ,就可能给出一个初值集合 ,从而得到一个相应的预报结果的集合 。目前 ,国际上先进国家与地区如美国 、欧洲中期天气预报中心(简称 EC) 等 ,根据上述思想均在 1992 年先后建立了集合预报系统(Ensemble Prediction System ,简称 EPS) 。中国国家气象中心在 1997 年 4 月也建立了 EPS ,每隔 5 d 利用谱模式 T63L16 作一次集合预报 ,初始扰动采用 3 d 的时间滞后法(LAF) ,共 12 个 10 d 预报成员(Member) ;继而在 1999 年夏天采用奇异向量法(SV) 生成初值扰动 ,共 32 个成员 ,利用谱模式 T106L19 作 32 个 10 d 的准业务预报 。其预报产品有 10 d 的 500 hPa 等压面上的集合平均高度场预报 ,850 hPa 等压面上的集合平均温度场预报 ,10 d 的 5 个等级(1 mm ,10 mm ,25 mm ,50 mm ,100 mm) 降水概率分布预报 ,4 个等级(- 8 K , - 4 K ,4 K ,8 K) 的 850 hPa 等压面上温度距平概率分布 10 d 预报 ,以及 850 hPa 等压面上集合平均的风向风速场的 10 d 预报 。

集合预报与传统的初边值问题的单一性的确定性预报不同 ,它是不确定性预报 ,从一群数据中提取有用信息 ,因此如何检验和评价集合预报至今还在研究探讨 ,下面就目前国际采用的集合预报检验方法 ,如 Talagrand 分布 ,离散度 、BS 评分 、信息检测理论应用于集合预报的检验 。

* 2000-11-22 收到 ,2001-04-16 收到修改稿 。

1 Talagrand 分布

O. Talagrand[1]认为一个“好”的 EPS 的标准应是每个预报成员似乎以同样的概率发生;换言之,观测实况也应以相同的概率落在它们附近。设 EPS 有 N 个成员,受检验范围为亚洲地区(60.25° ~ 146.25° E,20.25° ~ 56.25° N)。格点分辨率为 1.125 × 1.125 经纬度,则共有 $K = 2541$ 个格点,在每个格点 $j(j = 1, 2, 3, \cdots, K)$ 上,某气象变量的预报值可表示为 x_{ij}^f,其相应的观测值为 x_{nj}^o,其中 f 表示预报,o 表示观测,$i = 1, 2, 3, \cdots, N$ 表示每个成员的表示数。$n = 1, 2, 3, \cdots, L$,L 为实况场个数。将每个成员的预报值 x_{ij}^f 按数值增加的顺序排列,可有

$$x_{1j}^f \leqslant x_{2j}^f \leqslant x_{3j}^f \leqslant \cdots \leqslant x_{Nj}^f$$

$$d_i = x_{i+1,j}^f - x_{ij}^f \geqslant 0 \qquad i = 1, 2, 3, \cdots, N-1$$

d_o 为 $< x_{1,j}^f$ 最小端值外的区间

d_N 为 $> x_{N,j}^f$ 最大端值外的区间

这样按照 O. Talagrand 思想,观测值 x_j^o 必定落在某个区间 d_i 内,这里 $i = 0, 1, 2, \cdots, N$。随着用于验证的历史观测资料的增多,设 $n = L$ 则有效样本大小为 $M = LK$,观测值落在 $(N+1)$ 个区间 d_i 中的频数为 S_i,$i = 1, 2, 3, \cdots, N+1$,而它的期望值为 $M/(N+1)$ 则可求得频数 S_i 相对于期望值的均方差和观测值落在集合预报值的概率分布及概率均方差

$$D = \frac{1}{L}\sum_{j=1}^{L}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(S_{ij} - \frac{M}{N+1}\right)^2\right]^{1/2} \tag{1}$$

$$P_i = S_i / M \tag{2}$$

$$Q = \left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\overline{P} - P_i\right)^2\right]^{1/2} \tag{3}$$

其中平均概率 $\overline{P} = 0.03$,由概率分布 P_i 可绘得 Talagrand 分布图或称直方图,图 1 是国家气象中心在 2000 年 6 ~ 10 月利用奇异向量法产生初值扰动的全球谱模式 T106L19 对 500 hPa 亚洲地区高度场第 6 天集合预报进行检验的 Talagrand 分布,其中 $L = 85$,它的

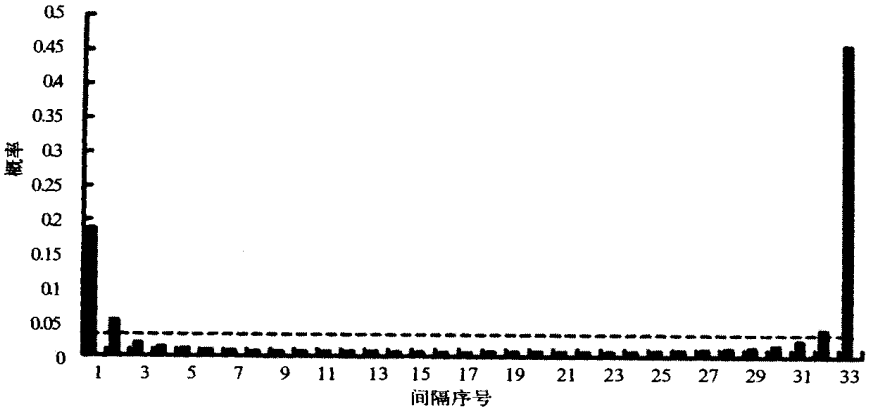

图 1 2000 年 6 ~ 10 月 H_{500} 的 Talageand 分布图(SV)(虚线为平均概率)

样本大小为 215985 。可见其概率分布基本是扁平的,但其两端点值为大,与在理想情况下的平均概率 0 .03(图中以水平虚线表示) 相比还是存在差距的,度量集合预报的优劣的另二个标准是 D 和 Q 值,它的第 6 天预报值分别为 209 和 0 .082 , D 和 Q 值越小越好,它表明了 EPS 更完美和有更高的可信度。为此,我们在 2000 年 9 ~ 10 月对 LAF 和 SV 初值扰动均利用 T106L19 谱模式制作集合预报进行了比较检验(为了对比试验 LAF 也采用 T106L19 谱模式) 。二者 25 个样本的 Talagrand 直方图与图 1 相似(图略) 。但 SV 的 $D = 222$, $Q = 0 .088$ 而 LAF 的 $D = 236$, $Q = 0 .132$ 。这说明 SV 法的集合预报比 LAF 的集合预报的可信度要高。

2　离散度

从日常众多的集合预报业务天气图上看,如 500 hPa 等压面上的高度场等值线图,会发现集合预报各成员的结果是离散的,说明了未来模式大气的预报不确定性。按照 R. Buizza[2]定义的离散度,可以看作为各扰动预报与控制 (未扰动) 预报之间的平均距离。设有 N 个成员的集合预报 f_i ($i = 1 ,2 ,3 , \cdots , N$) ,则集合预报的离散度 (S) 可用下式计算

$$S(t) = \overline{[\frac{1}{N}\sum_{i=1}^{N} d_i^2]^{\frac{1}{2}}} \tag{4}$$

其中 ——表示对 320 × 80 北半球高斯格点数值的平均, t 表示预报时效, $d_i = f_i(t) - f_0(t)$, f_1 为控制预报, f_i 为扰动预报 ($i = 2 ,3 , \cdots , N$) , f_o 为参考场,可为 f_1 ; f_o 也可看作为集合预报的平均

$$\overline{f} = \frac{1}{N}\sum_{i=1}^{N} f_i$$

若 f_o 是相应于预报的观测值,这就是集合预报平均的均方根误差,表 1 中 A ,B 列分别表示 2000 年 7 月 10 日 10 d 预报北半球相对于控制预报和集合预报平均的离散度。

表 1　H_{500}各预报时效 (h) 平均离散度　　m

	24	48	72	96	120	144	168	192	216	240
A	11 .7	16 .4	20 .7	19 .9	26 .4	40 .4	59 .6	44 .7	50 .0	61 .3
B	11 .5	12 .2	14 .4	17 .4	21 .0	25 .8	32 .2	36 .8	40 .0	40 .2
C	- 0 .79	- 6 .77	- 2 .2	- 2 .6	- 4 .7	- 5 .5	- 6 .4	- 6 .0	- 8 .9	- 9 .1
SV	22 .3	38 .9	53 .1	68 .0	82 .6	94 .1	102 .3	104 .7	106 .3	106 .1
LAF	59 .2	63 .6	71 .1	83 .2	93 .4	99 .6	106 .9	110 .3	109 .9	107 .2

可见集合预报的离散度要小些,一般来讲离散度小,可预报性大,集合平均预报要比控制预报可信度高。C 列表示 2000 年 6 ~ 8 月夏季 60 个样本的北半球集合预报平均的均方根误差与传统确定性预报(即控制预报) 的均方根误差之间的差值,其值均为负值,它说明了不确定性的集合预报平均结果比确定性的预报结果要好。因此,在大气稳定情况下,集合预报平均提供了未来大气的一种较好的可能性;但是在大气不稳定而出现分叉的多平衡态情况下,集合预报平均将失去意义,下文给出实例。而最后两列,表明了 2000 年 9 ~ 10 月 25 个样本的利用 SV 法和 LAF 制作 T106L19 全球 10 d 预报的平均均方根误差

比较,可见在 10 d 中 SV 的均方根误差均小于 LAF 的。因此可推知 SV 法的集合预报要比 LAF 更为准确。

离散度还可作为集合预报分群(Clustering)的量度。我们采用槽脊地理位置法将 2000 年 7 月 10 日 12:00 UTC 的第 6 天 32 个成员的 500 hPa 等压面上的高度场分成两组,再各自取该两组高度场的平均,第 1 组由 29 个成员组成,第 2 组由 3 个成员组成,然后在亚洲范围内计算它们的离散度,得 $S_1 = 2.35$ m, $S_2 = 219.37$ m。离散度小的第 1 组 500 hPa 高度场图与集合平均高度场图很相似,而离散度大的第 2 组 500 hPa 高度场与集合平均高度场有明显的差别。经实况检验,第 2 组的高度场更接近于实况,少数成员的平均图反而更正确,这是一个意外。因此离散度的大小可以作为集合预报分群的指标。其值大,有必要将集合预报分群,以向预报员提供更多的数值预报信息。

3 集合预报系统的概率预报检验

集合预报系统也是一种概率预报系统(PPS)。PPS 并不预报某种条件下的大气状态,而是预报大气状态(事件)发生的概率分布。目前的天气预报中早已存在了这种事件发生的概率预报,而且具有发展的趋势。我们按世界气象组织在 1989 年规定的方法[3]对集合预报系统中的概率预报产品进行检验。

3.1 Brier 评分

Brier(1950)定义了一种均方概率误差,称之为 Brier 评分(简称 BS):

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - O_i)^2 \tag{5}$$

其中 N 为二态分类事件的预报数;f_i 为事件发生的预报概率;如果事件发生 $O_i = 1$,事件不发生 $O_i = 0$。在这一形式中,评分的取值范围是 0 ~ 1,且越小越好,即 $BS = 0$ 表示概率预报最佳,预报正确;$BS = 1$,表示评分最差,预报失效。

经常使用的是 Brier 技巧评分(Brier Skill Score,简称 BSS),它是基于 BS 定义的,其表达式为

$$BSS = 1 - BS/BS_{clim} \tag{6}$$

其中 BS_{clim}为气候 BS 评分,且 $BS_{clim} = \overline{O}(1 - \overline{O})$,$\overline{O}$ 事件发生的气候频率

$$\overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i \tag{7}$$

BSS 表示了预报对气候预报改进的程度,若 BS 评分为气候值,则 BSS = 0。因此,若某事件的概率预报的 BSS > 0,则它的预报才有意义;反之 BSS < 0,则该事件的概率预报不如气候预测。由于采用了 BS_{clim},故 BSS 必须在足够大的样本中计算;否则,评分有很大的波动。越是少发生事件,越要求有较大的样本,以使评分稳定。与 BS 相反,BSS 值越大预报就越好,图 2 表示了 2000 年 6 ~ 10 月降水大于 1 mm 的 T106 模式 32 个成员集合平均的降水预报概率 76 个例子平均的 BS 和 BSS 评分结果,从 BS 曲线上看,BS 随着预报时间的延长而增大,即有无降水的落区预报准确率随着预报时效的延长也越来越差;从 BSS 曲线上看,其值均为负,它不如气候预测,参考价值小,但 BSS 随着预报时间延长而

减少,其预报准确率也随着预报时效延长而越来越差。图 2 中的三角形代表欧洲中期天气预报中心在欧洲地区从 1996 年 3 月到 1996 年 5 月的预报降水量大于 1 mm/ d 的10 d BSS 评分。其值为正值,具有参考价值。虽然用不同的年份和季节的 BSS 值来比较,由于资料有限和我国的 EPS 检验开发得晚,我们还可以看到 EC 的预报结果远比中国国家气象中心的为好。

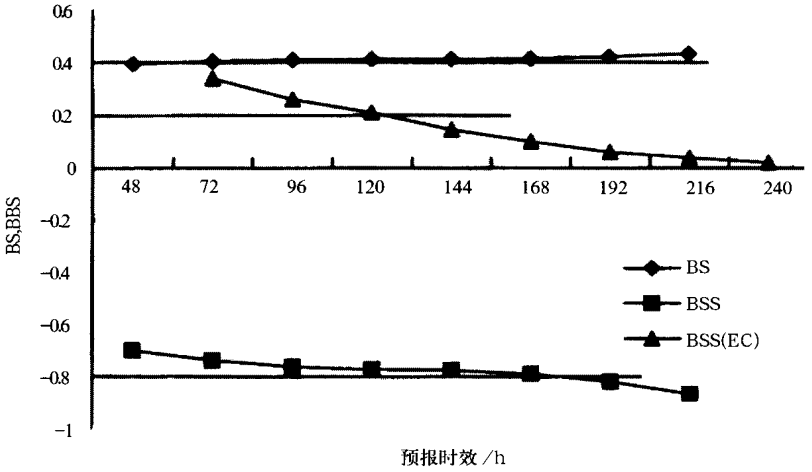

图 2　2000 年 6 ~ 10 月降水大于 1 mm 的 BS/ BSS 评分

而图 3 是对 2000 年 9 ~ 10 月 23 个样本的降水大于 1 mm 的 SV 和 LAF 法的概率预报 BS 和 BSS 检验比较,它表明了在 10 d 内 LAF 法的集合预报降水优于 SV 法的集合预报降水。

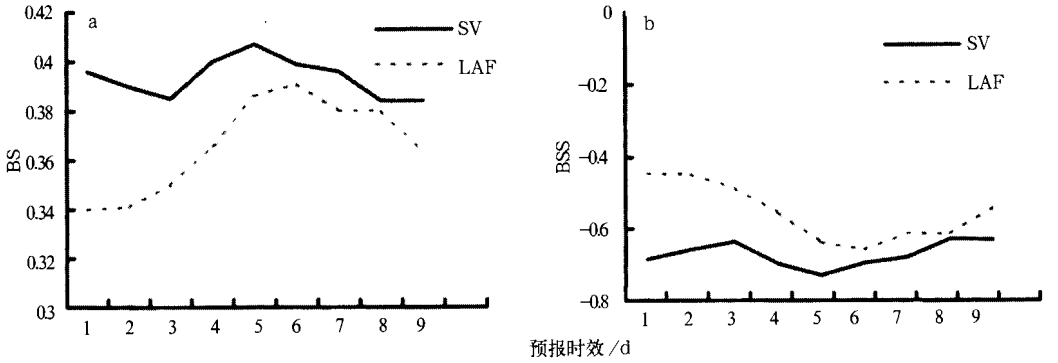

图 3　2000 年 9 ~ 10 月降水大于 1 mm 的 SV 和 LAF 法 BS(a) 和 BSS 评分比较(b)

3 .2　相对作用特征

相对作用特征(Relative Operating Characteristic,简称 ROC) 是信号探测理论(Signal Detection Theory) 在数值天气预报中的一种应用,对二分类要素序列进行检验。在每个格点上,考虑一个事件(如降水) 发生或不发生两种状态。用实况去检验预报,其结果必是下列情况之一,预报准确、正确否定、漏报和空报。这样一来我们可以构成如下的双态分类联列表(表 2)

表 2 双态分类联列表

	预报出现	预报不出现	观测相加
观测出现	X	Y	$X+Y$
观测不出现	Z	W	$Z+W$
预报相加	$X+Z$	$Y+W$	

其中 X 表示预报准确,用 f 表示命中率;Y 为漏报;Z 为空报,用 g 表示假警报率;W 为正确否定。信号探测理论主要使用上表中的两个量。

$$f = X/(X+Y) \tag{8}$$

$$g = Z/(Z+W) \tag{9}$$

由命中率和假警报率可在笛卡尔坐标上绘成一曲线,这曲线称之为 ROC 曲线。

下面将叙述如何把信号探测理论应用到集合预报与概率预报 ROC 检验上。设国家气象中心 EPS 的 T106 谱模式的 32 个成员在 2000 年 7 月 16 日 20:00(北京时)作出降水大于 1 mm 第 12~36 h 时效内的 24 h 的概率预报分布图,则 ROC 检验步骤如下:

(1)制作事件发生与否的统计表

在每个格点上,将预报概率分层为以 10%为宽度的纵向概率层次,对每个宽度统计事件发生和不发生的次数,制作出表 3。

表 3 降水观测统计(2000 年 7 月 16 日)

级序	概率范围(%)	未出现次数	出现次数
1	0~9	$a_1=71$	$b_1=149$
2	10~19	$a_2=20$	$b_2=18$
3	20~29	$a_3=17$	$b_3=14$
4	30~39	$a_4=11$	$b_4=11$
5	40~49	$a_5=14$	$b_5=8$
6	50~59	$a_6=9$	$b_6=9$
7	60~69	$a_7=12$	$b_7=10$
8	70~79	$a_8=15$	$b_8=14$
9	80~89	$a_9=27$	$b_9=14$
10	90~99	$a_{10}=39$	$b_{10}=39$

(2)计算命中率和假警报率

假定 30%为预报事件发生的概率临界值($i=3$,在表中画一水平虚线以示助解)。即预报概率大于 30%,则预报事件发生。给出这一临界值后,可以用求和的方式去得到表 2 中的 4 个量:

$$X_i=\sum_{j=i+1}^{k} b_j,\quad Y_i=\sum_{j=1}^{i} b_j,\quad Z_i=\sum_{j=i+1}^{k} a_j,\quad W_i=\sum_{j=1}^{i} a_j \tag{10}$$

其中 $k=10$,X_i 为水平虚线下出现次数的总和,Y_i 为水平虚线以上出现次数的总和,Z_i 为水平虚线下未出现次数的总和,W_i 为水平虚线以上未出现次数的总和。X_i+Y_i 为表中右列所有次数的总和,在表的下边给出;Z_i+W_i 是表中左列所有次数的总和。在 $i=3$ 情况下,利用式(8)和式(9)可求出命中率 $f=0.37$,假警报率 $g=0.54$。

通过在表上“移动水平虚线”可得不同的概率临界值,从而可计算得到其他的命中率和假警报率数据,如表 4 所示。

表 4　2000 年 7 月 16 日降水预报命中率和假警报率

预报概率(%)	0 ~ 9	10 ~ 19	20 ~ 29	30 ~ 39	40 ~ 49	50 ~ 59	60 ~ 69	70 ~ 79	80 ~ 89	90 ~ 99
命中率	0 .48	0 .42	0 .37	0 .33	0 .30	0 .27	0 .23	0 .19	0 .14	0 .00
假警报率	0 .70	0 .61	0 .54	0 .49	0 .43	0 .40	0 .34	0 .28	0 .17	0 .00

(3) 计算 ROC 面积

将命中率(f) 沿假警报率(g) 增加方向(x) 积分 ,就可得到 ROC 面积 。

$$\text{ROC 面积} = \int_0^1 f(x)\,\mathrm{d}x = \sum_{i=1}^{10} \frac{1}{2}[f_{i+1} + f_i] \times [g_{i+1} - g_i] \tag{11}$$

在上述例子中 ROC 面积 = 0 .168 。

如果检验量是降水 ,预报概率还须插入到观测站上 ,或者把降水插值到网格点上 。这里我们采用前者 。首先 ,在中国范围每天记录下所有降水发报站 ,并称之为检验站 。站上有降水时 ,概率为 1 ,无降水时则为零 。在 T106 模式情况下 ,对每一个测站均有 4 个格点将其包围(测站落在 1 .125 × 1 .125 个经纬度范围内) 。采用双线性内插就可得到该测站上的预报概率(或预报的降水量) 。在某天的 24 h 降水预报的实况出来后 ,就可按照上述步骤(1) ~ (3) 进行每日计算 。

图 4 就是 2000 年 9 ~ 10 月利用 LAF 和 SV 法制作的 T106L19 谱模式对降水大于 1 mm的预报时效 12 ~ 36 h 及 108 ~ 132 h 的 23 个例子的平均 ROC 曲线的比较 。它表明了事件预报的命中率和假警报率的相对比较的关系 。如 ROC 曲线越靠近图的左上方 ,命中率高而假警报率低 ,预报越好 ;反之亦然 。由图可见 ,SV 法在短期降水集合平均预报与 LAF 法差不多 ,而在中期第 5 天的降水集合预报比 LAF 法好 ,但它们的 ROC 面积均小于 0 .5 ,它们的预报已成为无技巧了 ,不能分辨出事件的发生与否 。

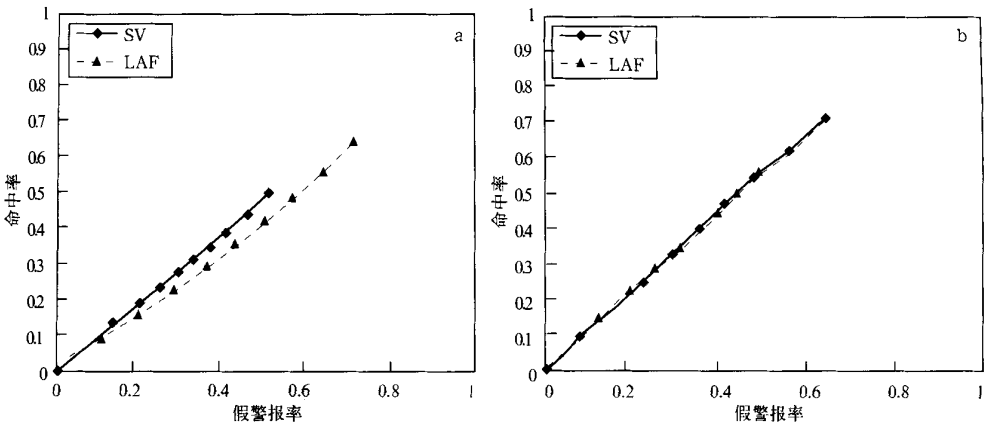

图 4　2000 年 9 ~ 10 月降水大于 1 mm 相对特征比较

(a) 108 ~ 132 h (b)12 ~ 36 h

4　结　论

本文简介了国家气象中心集合预报业务简况和集合预报产品及其检验方法 。并且对

2000 年 6 ~ 10 月的集合预报进行了统计检验,初步得到如下结果:

(1) 500 hPa 亚洲地区高度场第 6 天的集合预报的 Talagrand 分布直方图除两端外,其分布还是比较均匀的,离频数的期望值或概率均值还有一定的差别。

(2) 相对于集合预报平均场的离散度要比相对于控制预报(确定性预报)的离散度要小,即集合预报的可信度高。并且集合预报平均要优于日常的单一初值的控制预报。

(3) 离散度是集合预报分群的量度。离散度大,有必要将集合预报分成若干组群,以向预报员提供更多的气象预报信息。

(4) 通过降水概率预报的 BS、BSS、命中率和空报率的统计检验,目前的国家气象中心的降水预报还未达到作为业务参考的水平,且远比 EC 为差。

(5) 奇异向量法初值扰动的高度场集合预报要优于时间滞后法的集合预报。但是它的降水预报从 BB 和 BSS 上看不如时间滞后法。但中期(第 5 天后)降水预报从 ROC 曲线上看,SV 法又优于 LAF 法。

致谢:向作出 LAF 法 T106 集合预报试验的杨学胜、陈谊、应祝明表示感谢。

参 考 文 献

1 Talagrand O, Vautard R. Evaluation of probabilistic prediction systems. Workshop on Predictability ECMWF, 1997-10.

2 Buizza, Palmer T N. Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, 1998, **126**(9): 2503 ~ 2518.

3 Stanski H R, Wilson L J, Burrows W R. Survey of common verification methods in meteorology. WMO/TD-No. 358, 1989.

THE VERIFICATION FOR ENSEMBLE PREDICTION SYSTEM OF NATIONAL METEOROLOGICAL CENTER

Huangfu Xueguan

(*National Meteorological Center, Beijing* 100081)

Abstract

The various diagnostics, e.g., Talagrand probability distribution, spread, Brier score (BS), Brier Skill score (BSS), hit rate and false alarm rate are applied to the Ensemble Prediction System (the singular vector version, 32 members) at the National Meteorological Center, Beijing, China from June to October in 2000. The comparison also is conducted between Singular Vector (SV) and Lagged Averaging Forecast (LAF) methods.

Key words: Ensemble prediction Verification Comparison