

北半球 500 百帕组合因子及其在洞庭湖、鄱阳湖地区 7 月降水量预报中的应用

吴 贤 坂

(长江流域规划办公室)

提 要

用多元统计方法作水文气象长期预报,主要问题在于预报因子的选择。本文采用新的方法,选取高相关的预报因子——组合因子,并对其作若干性能的分析,经应用,收到了较好的预报效果。

一、前 言

多年来的预报实践表明,用多元统计方法作水文气象要素的预报,主要问题在于预报因子的选择。无论其方法在理论方面多么完善,在计算方面多么复杂,在功能方面多么全面,若没有一组具有一定质量的、一定物理意义的预报因子,仍将不能取得预定的预报效果。关于这一点,凡统计预报工作者多有共论。

为此,人们从多方面入手,去寻求所需要的预报因子。例如,天文因子,海洋因子,大气因子,地球物理因子,水文气象因子等。由于相关系数所能达到信度的高低,大体上能标志两个要素之间的密切程度;而且,相关程度越高,预报效果也就相应地比较好。因此,目前多采用相关普查,结合物理意义的分析,选取符合一度信度相关的预报因子。

本文将针对北半球 500 百帕高度场资料和长江流域水文气象要素,采用一种新的相关普查方法,选取一种高相关的预报因子——作者称之为组合因子,并试作其若干性能的分析。实践表明,多元统计方法中使用这种组合因子,能得到较好的预报效果。

二、组合因子的计算

(一) 思路

目前,从北半球 500 百帕月平均高度场选取的预报因子,大体有关键区要素值、西风风速、环流指数和大气活动中心位置等。

选取关键区要素值的一般作法是:计算预报要素与高度场每个网格点的单相关系数;确定符合某信度的相关系数值所对应的网格点,如果此类网格点连成一片,则把这一

片作为关键区;分析该关键区的物理意义;如果决定选取,则计算关键区内各网格点的逐年高度和作为预报因子。但是,如何划分此类关键区的边界,往往因人而异,特别是当连成一片且符合某信度的网格点比较多或少时,更是如此。另外,关键区要素值同预报要素的相关系数往往低于单网格点的相关系数。

作者考虑到上述情况,并认为一个关键区的变化一般不能全面反映环流变化的特征。由此设想,如果能冲破关键区的边界,在整个北半球挑选符合某信度的相关系数值所对应的网格点,组成一个不连成一片的“关键区”,以其各网格点的线性组合作为预报因子,使它同预报要素的相关系数值达到最大,其代表性可能比较全面些。这就是本文标题中所说的“组合因子”的概念。

(二) 计算方法

根据上述思路,可以把组合因子的计算归纳为以下7个方面。

1. 以公式

$$R_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (1)$$

计算预报要素 Y 与北半球 500 百帕月平均高度场网格点 X 的单相关系数。其中 $\bar{X}$ 和 $\bar{Y}$ 分别为 X 和 Y 的平均值, n 为样本容量。

2. 挑选符合给定信度的相关系数值所对应的网格点,记为 $X_i (i = 0, 1, 2, \dots, m)$, 并把绝对值最大的相关系数值所对应的网格点记为 X_0 。

3. 以 X_0 的相关系数 $R_{X_0Y}^{(0)}$ 的符号(正或负)为准,如果 $X_i (i = 1, 2, \dots, m)$ 各点中,有某些点的 $R_{X_iY}^{(0)}$ 的符号与 $R_{X_0Y}^{(0)}$ 的符号不一致,则改变这些点高度值的符号,从而改变了这些点的相关系数的符号。这是为了达到把相关场上主要的正、负相关网格点都吸收进组合因子的目的。

4. 以公式

$$H_i^{(j)} = H_0^{(j)} + H_i^{(0)} \quad (j = 0, 1, \dots, m; i = 1, 2, \dots, n) \quad (2)$$

计算高度和。其中 $H_0^{(j)}$ 和 $H_i^{(0)}$ 分别为 X_0 和 X_i 各点的原始高度值。再用公式(1)计算 $H_i^{(j)} (j = 0, 1, \dots, m; i = 1, 2, \dots, n)$ 与 Y 的相关系数 $R_{X_iY}^{(j)}$, 并选出其中绝对值最大的相关系数 $R_{X_KY}^{(1)}$ 及其对应的网格点 $X_K (0 \leq K \leq m)$ 。

5. 判断下列不等式是否成立:

$$R_{X_KY}^{(1)} > R_{X_0Y}^{(0)} \quad (0 \leq K \leq m) \quad (3)$$

6. 如果(3)式成立,则重复第4—5点。不过,此时(2)式中的 $H_i^{(j)}$ 应为 $H_i^{(j)}$, $H_0^{(j)}$ 为 $H_K^{(j)}$ 所代, $H_i^{(0)}$ 为累加高度值 $H_i^{(j)}$ 所代替。如果(3)式不成立,则转第7点。

如果进行到第 l 次,(2)式和(3)式可改为

$$H_i^{(j)} = H_K^{(j-l)} + H_i^{(j-l)} \quad (2)'$$

$$R_{X_KY}^{(l)} > R_{X_0Y}^{(l-1)} \quad (3)'$$

其中 $j = 0, 1, \dots, m; i = 1, 2, \dots, n; 0 \leq K \leq m$ 。

7. 选取上述计算所得线性组合 $HK_i^{-1}(i = 1, 2, \dots, n; 0 \leq K \leq m)$ 为 Y 的预报因子, 即所谓的组合因子。

三、组合因子实例分析

在本节中, 将以长江流域洞庭湖、鄱阳湖地区(以下简称两湖地区)7 月降水量为例, 说明组合因子的计算和形成过程, 并进行本文前言中所提到的几个方面的分析。

(一) 组合因子的形成

根据两湖地区 7 月降水量与北半球 500 百帕月平均高度资料的相关普查计算结果, 它们之间有较好的隔年相关。因此, 在本节中均以头年 8 月北半球 500 百帕月平均高度(记为 H_8) 与当年 7 月两湖地区降水量(记为 R_7) 的计算为例, 进行叙述。 H_8 的资料取自 1951—1981 年, R_7 的资料取自 1952—1982 年。

按照第二节所述的计算步骤进行计算, 得到一个组合因子。 它由 11 个网格点组成(信度为 5%), 其中有两个网格点是同一点, 故实际只涉及 10 个网格点。 它们进入组合因子的顺序、位置和相关系数列于表 1, 并把组合因子的相关系数随网格点数增加而变化

表 1 组合因子中各网格点的顺序、位置和相关系数

序 号	1	2	3	4	5	6	7	8	9	10	11
位置	50°N 50°E	50°N 120°W	35°N 125°W	60°N 40°W	50°N 100°E	45°N 115°W	60°N 50°E	55°N 45°W	30°N 130°W	35°N 45°E	50°N 100°E
初始相关 系数 $R^{(0)}$	0.58	-0.48	-0.58	-0.37	-0.40	-0.36	0.54	-0.40	-0.36	0.37	-0.40

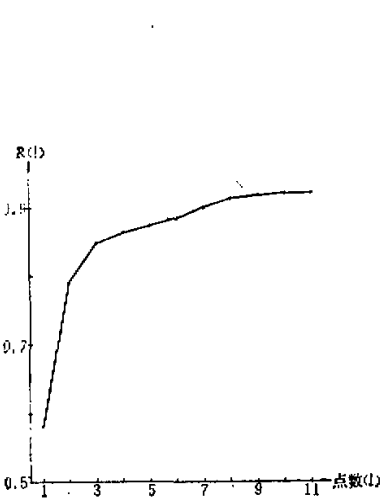


图 1 组合因子的相关系数随点数增加而变化

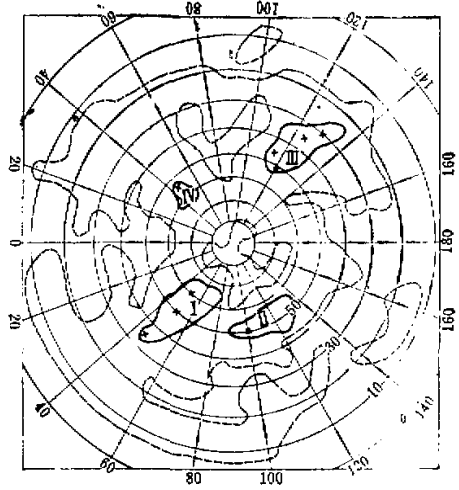


图 2 组合因子中各点的位置及其代表性
“×”为组合因子点的位置; 实线为关键区边界; 虚线为 5% 信度的相关系数等值线。

的过程绘于图1。表1中第1、3两点的初始相关系数分别为0.58和-0.58,是整个北半球相关场上绝对值最大的两个点。

在北半球500百帕相关场上,按5%信度共有4个较大的相关区,分别位于乌拉尔(I区)、东亚(II区)、太平洋东岸(III区)和大西洋北部(IV区),如图2中实线所围区域。组合因子中的网格点如图2中“×”号所在位置。可见组合因子汇集了4个较大相关区中的10个点。图1表明,组合因子的形成过程中,相关系数不断增大。第1、2两点组合时,其上升幅度最大;随着点数增加,上升幅度逐渐变小;8点以后,曲线趋于平缓。

(二) 拟合效果

把组合因子(记为 $H^{(1)}$)与两湖地区7月降水量(记为 R_7)的过程线绘于图3。图3两

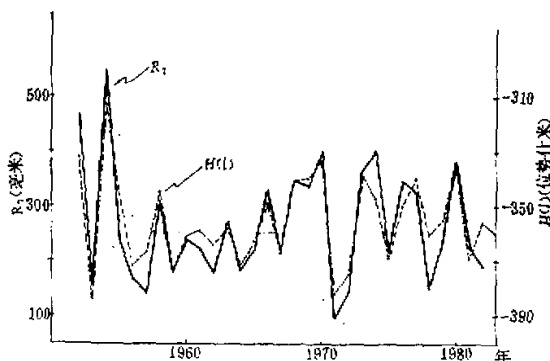


图3 两湖地区7月降水量(R_7)与组合因子($H^{(1)}$)逐年变化

曲线的吻合程度比较好,特别是在极值处能有这么好的对应,一般是不容易的。

为了使图3的拟合效果定量化,应用回归分析对 $H^{(1)}$ 和 R_7 计算,得一元回归方程:

$$R_7 = 2377.22 + 5.94H^{(1)} \quad (4)$$

把(4)式的计算值与 R_7 实测值的过程线对照于图4。比较图3和图4,可见差异很小。有关回归量的计算结果列于表2。表2中的各数据指标,大体上可以代表图3的相应水平。

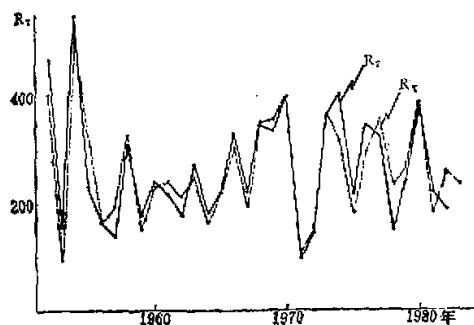


图4 两湖地区7月降水量(R_7)与(4)式计算值(R_7')对照

表 2 一元回归计算结果

拟 合 误 差						残差 平方和
最大	%	最小	%	平均	%	
87.6	37.0	0.8	0.2	31.3	12	42.1

(三) 代表性分析

组合因子的 10 个网格点是否能反映出 4 个关键区的变化,乃至整个北半球环流形势的变化,作者计算了它们与北半球各网格点的相关系数,其中相关系数值达各种信度的网格点数列于表 3。表 3 中右边第 1 列的百分率(%)按 $\frac{\text{净点数}}{576} \times 100\%$ 计算。表 3 表明组合因子中的每个网格点在整个北半球都占有一定的地位。组合因子的变化,能以 5% 的信度波及北半球 60% 区域的变化。

表 3 组合因子中的网格点与北半球各网格点的相关系数达某信度的网格点数

组合因子中的网格点序号	1	2	3	4	5	6	7	8	9	10	合 计		
											总点数	净点数	%
$ R \geq R_{0.991}$ 的点数	16	5	9	11	12	9	22	10	17	8	119	91	16
$R_{0.991} > R \geq R_{0.91}$ 的点数	25	11	12	11	9	16	7	5	23	15	134	100	17
$R_{0.91} > R \geq R_{0.95}$ 的点数	29	26	24	24	38	54	27	17	84	25	348	153	27
合 计	70	42	45	46	59	79	56	32	124	48	601	344	60

经计算,若把信度降为 10%,则这 10 个网格点的相关点数可达 450 个,占北半球的 80%,几乎覆盖着整个北半球。

把 10 个网格点的相关区域用 5% 信度相关系数等值线绘于图 2 (虚线),便可看出相关区域遍及东、西两半球,以及高、中、低纬度,足见组合因子的代表性。

(四) 稳定性分析

这里所说的稳定性是指 $H^{(t)}$ 与 R_t 之间的关系。随着时间的延续,其相关性是否能保持相对的稳定,作者用第二节中的计算方法,计算了 12 个组合因子。其中第 1 个组合因子的计算资料是 20 年($H^{(1)}$ 取 1951—1970 年, R_t 取 1952—1971 年);第 2 个组合因子的计算资料是 21 年($H^{(2)}$ 取 1951—1971 年, R_t 取 1952—1972 年);余类推。图 5 绘出了 12 个组合因子的相关系数值 r_i (实线)和 $r_i - R_{0.001}$ 值(虚线) ($i = 1, 2, \dots, 12$) 的过程线。其中 $R_{0.001}$ 为信度取 1% 的相关系数临界值。从图 5 中的实线看,相关系数值经过 12 年的检验,一直能保持在 0.90 以上,处于比较稳定的状态。



图 5 相关系数随时间变化过程线

从图 5 的虚线看, 差值随着时间的增加而逐渐增大。这表明相对于 1% 的信度来说, 随着时间的延续, 组合因子与预报要素的相关关系比较稳定, 并渐趋密切。

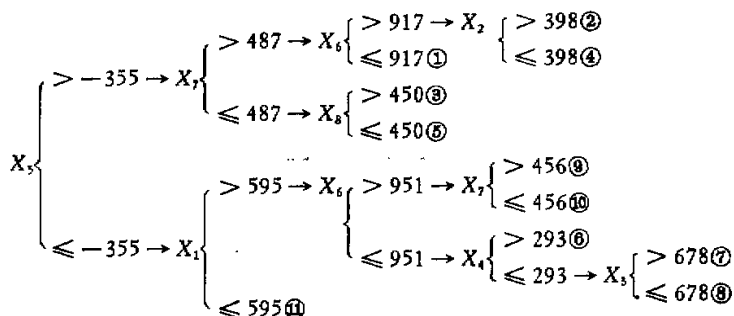
为了检验上述关系, 作者分别计算了 11 个组合因子和预报要素的距平值, 并把组合因子在预报年数据的距平符号和预报要素相应的实测值距平符号列于表 4。从表 4 看, 距平符号一致的年份有 8 年, 定性预报准确率可达 73%。由于上述预报与实测是相互独立的, 即组合因子和预报要素在预报年的数据都不参与相关计算, 自然就增加了这种检验所得结果的可靠性。

表 4 组合因子和预报要素在预报年的数据距平符号对照

预 报 年 份		1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
预报年数据 的距平符号	组合因子	-	+	+	-	-	+	-	+	+	-	+
	预报要素	-	+	+	-	+	+	-	-	+	-	-

(五) 组合因子的作用

组合因子同预报要素的相关系数数值比较大, 因此, 它的预报贡献也往往比较大。下面是预报要素 R_7 应用聚类分析中的 A. I. D 方法所作的一个预报方案。其中圆圈中的数字代表所在预报值的大小顺序。例如, ④所在的预报值最大, ⑩所在的预报值最小, 余类推。其中 ①—⑥ 所在的预报值均大于历年均值, ⑦—⑩所在的预报值均小于历年均值。



字代表所在预报值的大小顺序。例如, ④所在的预报值最大, ⑩所在的预报值最小, 余类推。其中 ①—⑥ 所在的预报值均大于历年均值, ⑦—⑩所在的预报值均小于历年均值。在预报方案中, X_5 是组合因子, 它同 R_7 的相关系数为 0.92。此方案说明, 经 X_5 的一次分类, 就能把 ①—⑥ 分到一个方向, 而另一个方向的预报值都是小于均值的。

不难验证, 在逐步回归计算中, X_5 的方差贡献在同一批预报因子中为最大, 并将首先进入回归方程。在实际预报应用中, 例如 1981 年长江上游出现涝的现象, 1983 年长江中下游出现涝的现象, 在有组合因子参加的预报方案中, 都能比较好地反映出这种趋势。

(六) 组合因子与回归方程

从组合因子的形成来看, 它与回归方程是有区别的。例如:

1. 引进原则

组合因子中各网格点的进入, 是以相关系数绝对值增大最快为原则的; 回归方程中各变量的引进, 是以其方差贡献最大, 从而使残差平方和下降最大为原则的。

2. 引进顺序

表 5 组合因子和逐步回归引进网格点的顺序及其系数

网格点序号		1	2	3	4	5	6	7	8	9	10	11
组合因子	顺序	1	2	3	4	5	6	7	8	9	10	11
	系数	+1	-1	-1	-1	-1	-1	+1	-1	-1	+1	-1
逐步回归	顺序	1	2	3	4	7	6	5	8	9	10	—
	系数	6.73	-8.19	1.17	-5.50	6.39	-7.42	-8.78	-6.66	-16.62	4.75	—

由于引进原则不同,使得引进的顺序也不一致。作者仅对组合因子中的 11 个网格点应用逐步回归分析进行筛选(取 $F = 0$),并把两种方法引进网格点的顺序和系数列于表 5。表 5 表明,前面 4 点的顺序一致,从第 5 网格点开始,顺序就不一样了。由于第 11 点与第 5 点是同一网格点,在逐步回归分析中自然不会被引进回归方程。

3. 系数

组合因子中各网格点的系数都是整数,除了第 5 和第 11 点是同一网格点,其系数为 2 以外,其余系数都是 1;回归方程中各网格点的系数则不同。两者的系数符号亦略有区别。

4. 显著性

组合因子的显著性可用最终相关系数来表示,为 0.92;回归方程的显著性可用复相关系数来表示,经计算为 0.93。可见后者的拟合率比前者要高。把逐步回归的计算结果列于表 6,并与表 2 比较,其拟合差异是显而易见的。

表 6 逐步回归计算结果

拟合误差						残差平方和
最大	%	最小	%	平均	%	
101.8	43.7	0.5	0.4	29.0	11.0	47.8

四、问题讨论

1. 本文是作者近年工作的一个小结,故具有一定的经验性。已经表明,计算北半球 500 百帕月平均高度场的组合因子是可行的。同样,这种方法也适用于其它资料场的计算。例如海平面气压场、海表温度场、水文气象要素场等。凡具有一定数量的资料群,计算的相关场上有 m 个相关系数绝对值大于给定信度的相关系数临界值,则所得组合因子的最终相关系数必将大于原始相关系数。

2. 第二节第(二)段第 3 点的计算是必要的。否则,当 $R_{x_0y}^{(0)} > 0$ 时,组合因子将把 $R_{xy} < 0$ 的那些高相关网格点拒之在外;当 $R_{x_0y}^{(0)} < 0$ 时,组合因子将把 $R_{xy} > 0$ 的那些高相关网格点拒之于外。从而,降低组合因子的功能,失去它的代表性。

3. 每个组合因子与预报要素的相关系数值并不都是很大的。作者针对两湖地区 7 月降水量与前期北半球 500 百帕月平均高度场,从头年 1 月至当年 4 月都进行了计算,各组

合因子的相关系数过程线如图 6。对曲线作 2 点滑动平均,如图 6 中虚线,趋势更加明显。

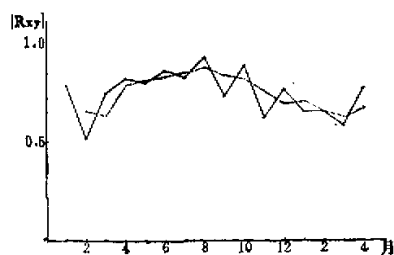


图 6 各月组合因子的相关系数绝对值过程线
实线为 $|R_{xy}|$, 虚线为 $|R_{xy}|$ 的两点移动平均。

这意味着,前期北半球 500 百帕月平均环流对两湖地区 7 月降水量的影响,从头年 4 月开始增强,至 8 月最为显著,然后逐月衰减,至当年 3 月最弱,4 月开始又转为增强。这种现象可能有助于探索长期天气过程的变化。

4. 将组合因子用于多元分析方法做预报时,与其它因子配合使用,效果较好。作者作了 1982 年的预报试验,结果列于表 7。表 7

中的 1 号试验,参加计算的 8 个预报因子都是组合因子;3 号试验中,参加计算的 8 个因子都是非组合因子;2 号试验中,参加计算的 8 个因子为 1 号试验中的 4 个组合因子和 3 号试验中的 4 个非组合因子。从表 7 可以看出,2 号和 1 号试验的预报效果比较好。

表 7 各种预报因子的预报效果试验(预报和实测值的单位:毫米)

项 目	试验序号		1		2		3	
	预报方法		逐步回归	A. I. D	逐步回归	A. I. D	逐步回归	A. I. D
预报值			231	230	214	210	238	274
实测值			192 (历年均值为 267)					
误差%			20.3	19.8	11.5	9.4	24.0	42.7

表 8 三种试验中逐步回归计算结果比较

回归量 试验号	F 水平	残差平方和	因子个数	复相关系数	预报值	误差%
1	2.5	30.8	3	0.94	231	20.3
2	2.5	28.2	4	0.96	214	11.5
3	2.5	43.1	5	0.90	238	24.0

进一步把表 7 中逐步回归计算的几个量列于表 8。比较起来,2 号试验的回归效果最好,3 号的最差。表 7 和表 8 还反映出,当组合因子与其他非组合因子配合使用时,预报效果比较好些。

500 hPa COMBINATION FACTORS ON NORTHERN HEMISPHERE AND ITS APPLICATION TO THE FORECASTING OF JULY PRECIPITATION IN THE REGIONS OF DONGTING AND POYANG LAKES

Wu Xianban

(Changjiang Valley Planning Office)

Abstract

The key problem about applying the multiple statistical method to the long-range hydrometeorological forecast is the selection of forecast factors. In this paper a new approach is used to select the close-correlation forecast factor, named as combination factor. Some characteristics of the combination factor are analysed. The more effective result of forecast is achieved by applying this approach.