

黄河三花地区汛期逐日降水 MOS 预报的 因子选择试验*

黄 嘉 佑

(北京大学地球物理系, 100871)

符 长 锋

(河南省气象科研所)

河南黄河三花地区地处黄河流域的中游下段地区, 该地有名的黄泛区是历史上水旱灾害频发的地区。该地汛期短期降水预报的准确与否是十分重要的。由于近年来数值预报的发展和广泛应用, 局地气象要素短期预报大多使用 MOS 方法^[1], 其中降水预报常用一种所谓 REEP^[2]的降水概率预报。本文主要研究用此方法作该地汛期逐日降水 MOS 预报过程因子选择问题。

在降水概率的 MOS 预报中, 需要建立包含多个预报因子的预报方程, 需在众多的因子中选取有代表性的因子进入方程。本文将用逐步回归技术来选择满足这一要求的因子^[3]。

本文提出 4 种选择因子方案。它们是: 线性相关逐步回归, 非线性逐步回归, 主分量非线性逐步回归及转动主分量非线性逐步回归等 4 种方案。用它们对因子进行筛选, 试验结果表明转动非线性主分量逐步回归的方案所选择的因子具有最好的效果, 且有较强的天气意义。

1 资料与方法

表 1 各物理量场的意义

序 号	名 称
01	700hPa T-T _a 初始场
02	700hPa Qse初始场
03	700hPa 水汽通量初始场
04	700hPa 水汽通量散度初始场
05	500hPa 地转涡度初始场
06	700hPa 垂直速度初始场
07	500hPa 地转涡度36小时预报
08	700hPa 垂直速度36小时预报
09	700hPa 水汽通量36小时预报
10	700hPa 水汽通量散度36小时预报
11	700hPa T-T _a 36小时预报
12	700hPa Qse36小时预报
13	12—36小时降水量预报
14	24小时降水量预报

取国家气象局所发的细网格 MOS 输出物理量场 24 个, 资料年代为 1985 年至 1988 年。但台站保证能逐日收到的只有其中的 14 个场, 各场范围在我国黄河与长江流域之间的正方形范围内, 其中含 9 行 11 列的 99 个网格点(参见附图)。各物理量场的意义列在表 1。

我们取 1985—1988 年 7 月 15 日—8 月 13 日该地区各站逐日降水量的资料, 又取同期前一天 14 个 MOS 物理量场资料。然后逐日将该区所有站降水量求和, 由此得该区降水量序列。这一序列作为预报对象, 用上述各物理量场各网格点作为因子并建立 REEP 降水概率预报方程。预报量的 0, 1 值以平均值(4.3 mm)为分界值。

在建立预报方程中需对上述 14 个物理量场 1386 个因子进行筛选。一般因子的筛选是以因子与预报量的线性相关系数是否显著(显著水平

5%临界相关系数为 0.18)作为选择的标准。但由于气象要素之间关系的复杂性, 仅考虑它们之间的

* 1991 年 6 月 7 日收到原稿, 1991 年 8 月 3 日收到修改稿。

线性关系是不够的, 还应考虑其非线性关系^[4]。因此选择因子时, 应选取线性和非线性相关系数显著的因子。预报量 y 与因子 x 的线性和非线性相关关系常用的有线性、指数、自然指数和对数等 4 种类型, 函数形式可用最优化方法确定^[4]。然后把函数作为新变量用常用方法求出其各类型的非线性相关系数^[4]。

为预报需要, 除考虑预报量与因子的相关关系密切程度外, 还应考查因子的预报能力。这种预报能力与预报量和因子的关系稳定性有关。其关系稳定性可用它们的线性和非线性相关系数变化情况来看表现。设所取的样本容量为 N , 分别再取其中的 $N-1, N-2, \dots, N-m$ 样本容量计算它们之间相关系数 $r_1, r_2, \dots, r_{m+1}$ 。显然, 它们的变化系数可表为:

$$CV_r = S_r / \bar{r} \quad (1)$$

其中

$$S_r = \left[\sum_{i=1}^{m+1} (r_i - \bar{r})^2 / (m+1) \right]^{1/2}$$

$$\bar{r} = \sum_{i=1}^{m+1} r_i / (m+1)$$

并称为因子的预报稳定性。当某一因子与预报量的相关系数的变化系数有较小值且它与预报量有较高的相关系数时, 则这一因子为预报能力较强的因子。计算时(本文取 $m=5$)比较不同类型因子预报能力, 取其中最强者为入选因子。我们用这种方法对 14 个物理量场上的因子进行筛选。待选因子群可以是各物理量场中单个格点, 也可以是由各场格点综合而成的新因子。

2 网格点上逐步回归的因子筛选

为了克服对所有物理量场作考虑时因子数量太大而产生的困难, 我们采取对逐个物理量场上的因子用逐步回归方法^[3]进行筛选。这种方法可保证所选择的因子既与预报量有密切关系, 它们之间关系又较小, 从而能作为该物理量场的代表性好的因子(对预报量而言)。又考虑到实际因子与预报量的显著相关系数个数不是很多, 我们选取 15 个预报能力较强的因子作为粗选, 建立每个场的待选因子群。然后用逐步回归进行第一次筛选。最后再把各场中的代表性好的因子集中在一起用逐步回归进行第二次筛选, 从而建立最后的降水概率预报方程。我们把仅用线性相关系数作上述筛选称为线性相关逐步回归方案(I), 而把用非线性相关系数作上述筛选称为非线性相关逐步回归方案(II)。用 $F=3.0$ 作逐步回归两方案比较。结果发现, 在同一选择标准下, 方案(II)选入的因子要比方案(I)多, 因子的相关系数(绝对值)要高, 方程的解释方差也要高。

3 物理量场的综合因子

考虑到单个网格点作为因子时仅能反映物理量场中该点物理因素的作用, 其天气意义是十分有限的。而且单个格点的序列随机性较大, 与预报量的关系不容易保持稳定。因此寻找物理量场的综合因子来代替单个格点因子是十分必要的。我们考虑使用主分量分析方法^[5]提取各个物理量场的主要分量作为场的综合因子, 这些因子能反映场的主要变化特征, 而且它们之间是相互正交, 完全能满足场的代表性要求。14 个物理量场前 10 个综合因子所反映场的方差百分率可达 50% 以上。

我们对 14 个物理量场的 140 个综合因子作与上述筛选方法类似的逐步回归建立预报方程, 并称为主分量逐步回归筛选方案(III), 其结果列在表 2。从表 2 中可见, 所选入的综合因子均为非线性因子, 反映物理量场与预报量关系的复杂性。但是, 用此方案所建立的方程对预报量的解释方差并不比单个网格点要高, 所有因子与预报量的相关系数并不很高。其原因可能是由于综合因子所反映, 是整个场范围内的特征, 而局地降水只与场中的某一地区有较大的关系。

表 2 主分量与主分量逐步回归筛选方案比较*

(III)						(IV)					
v	i	L	a	b	r	v	i	L	a	b	r
1	2	4		-.163	.201	5	6	4		-.225	.276
2	5	4		-.219	.231	6	7	2	1.119	-.243	.298
5	7	4		-.151	.200	8	3	1		.157	.297
6	7	4		-.181	.206	11	1	4		-.207	.330
8	8	2	1.166	-.201	.211	13	3	3	-3.953	-.324	.402
13	7	4		-.152	.204						
14	2	3	2.149	.261	.284						
复相关系数			.547						.616		

* v 表示物理量场序号, i 表示综合因子序号, L 表示因子与预报量关系函数类型, a 表示函数中的常数, b 表示方程中标准化变量回归系数, r 表示非线性相关系数

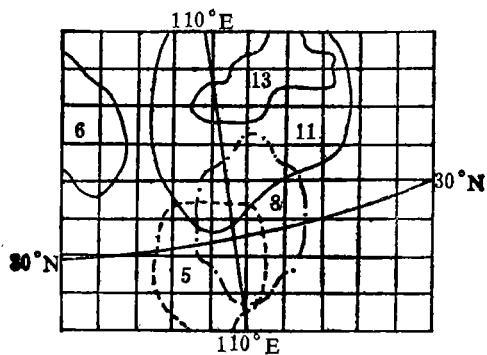


图 1 各因子高荷载区分布

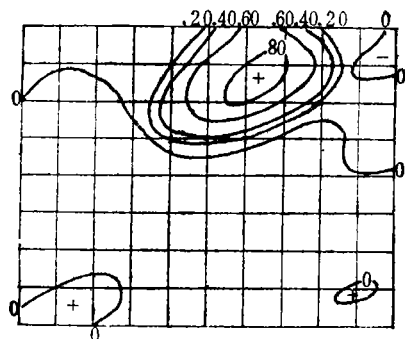
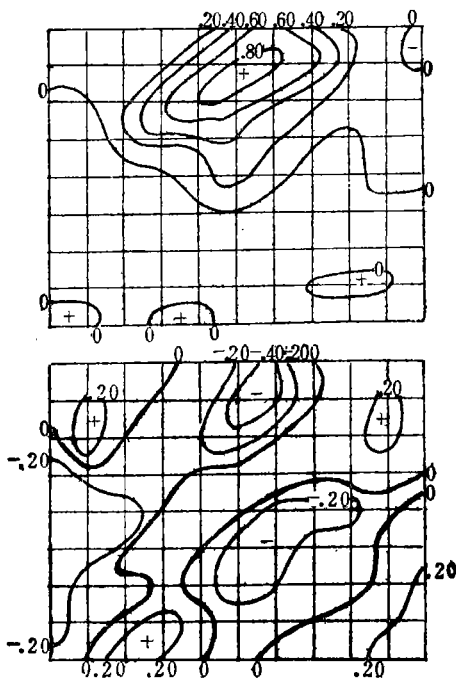
图 3 第 13 个物理量场相关场
(以格点 17 为基点)

图 2 第 13 个物理量场第 3 个转动主分量(上)与第 7 个主分量(下)的荷载场分布的比较

为了寻找场内局地范围的特征,并提取该范围综合因子,我们使用转动主分量分析方法(方法详见参考文献[5]和[6])对原主分量进行转动。考虑到在方案(III)中所选入的原综合因子仅在各物理场的前 8 个主分量中,而且除个别场(如 10, 13 和 14)外,其余各场前 8 个主分量的解释方差均可达 80% 以上。因此对各场前 8 个分量作最大正交转动。得到的转动主分量作为新的场综合因子。用它们作类似的筛选,并称为转动主分量逐步回归筛选方案(IV),其预报方程列在表 2 右边。与方案(III)比较可见,

所选入的因子相关系数均有所提高,其预报方程的解释方差比方案(III)有所提高。在所有方案中进入方程的因子个数来说它是最少的,说明其中的因子有较强的解释能力。

4 因子的天气学意义

由于方案(IV)有最好的因子筛选效果,我们主要讨论所选入回归方程中的因子的天气学意义。图1给出各因子高荷载区(即它与物理量场各点相关系数(绝对值)高于0.60的相关区)的分布情况,图中区域内的数字为物理量场序号。

表明黄河三花地区汛期逐日降水与前一天的地转涡度,垂直速度,降水等预报场有密切关系。它们分别处在所预报地区的西部,北部及西南部地区。说明这些地区的高层地转涡度和垂直速度对该地降水有重要影响,另外该地的湿度露点差和降水量的预报值也有密切关系。事实上这些地区也为单点线性(方案(I))和非线性(方案(II))回归所选择的因子出现较多的地区。

这些地区常常可以反映物理量场的某些主要特征。我们以第13物理量场为例作一说明。图2给出进入方程中方案(IV)第13个物理量场第3个转动主分量与方案(III)同一物理量场第7个主分量对应的荷载场的分布的比较。从中可见,前者高荷载量集中在北部中央部分,其转动主分量主要是反映该地区的变化特征。而后者则没有较高的荷载区,稍高的8个相关区(绝对值大于0.20)分散分布在各个方位,其对应的主分量仅反映这种分布的总体特征。而预报员经验表明该地降水常与天气形势场中的局地变化有密切关系。转动主分量由于在局部地区上有十分高的荷载,而在其它地区则表现为十分低的荷载。因此它能较好地反映局部地区的变化特征,这与预报员经验相符。为了进一步说明这一点,图3给出以该物理量场北部一格点为基点的相关图(即该点与场中其余各点的相关系数分布)。从图3可见,在场的北部为一片高正相关区,其中相关系数高于0.40所包围的地区与该场第3个转动主分量高荷载区十分一致。在这一区内格点19的序列在方案(I)和(II)中均被选入回归方程中,它与预报量的线性和非线性相关系数分别为0.281和0.353,而该场所选的转动因子与预报量的相关系数为0.352。说明转动因子具有较强的天气意义且有较强的可预报性。

5 预报稳定性与预报检验

我们把各方案所得的方程中所有因子的预报稳定性的平均值作为方程预报稳定性的度量。上述4种方案方程的稳定性分别为0.044,0.048,0.029和0.039,表明后两种方案有较高的稳定性。把它们用在独立样本(1989年7月15至31日)的17天进行预报检验,结果报对的天数分别为10,10,6和10。后两种方案并未表现出较高的预报准确率,其原因是由于在预报中主要分量和转动因子未来值的确定是假定荷载场不随时间变化情况下得到的。而实际上当资料增加时荷载场会发生变化的。为避免这一误差,在方案(IV)基础上,我们采取根据转动因子高荷载区求区域格点平均序列以代替相应的转动因子,再用上述方法建立回归方程。比较发现新序列与转动因子在与预报量的非线性相关系数的数值上无太大差别,个别甚至有所提高,说明它们可代替相应的转动因子。在独立样本预报检验结果(报对12天)表明预报效果有所提高。

6 结论与讨论

对黄河三花地区汛期(7月15日至8月15日)逐日降水多种MOS预报方案的试验表明转动非线性主分量逐步回归的方案所选择的因子具有最好的效果,且有较强的天气意义。在此方案基础上,进一步根据转动因子高荷载区求区域格点平均序列以代替相应的转动因子,在预报中具有更好的效果。

当然独立样本预报检验效果还需进一步提高,但这种提高也有赖于数值模式的改进。在这方面还有待作进一步的研究和探索。

参考文献

- [1] 黄嘉佑. MOS 方法的进展. 气象科技, 1984(3): 18—21.
- [2] Glahn H R, Lowry D A. The use of Model Output Statistics (MOS) in objective weather forecasting. J Appl Meteor, 1972(11): 1203—1211.
- [3] 黄嘉佑. 气象统计分析与预报方法. 气象出版社, 1990. 387.
- [4] 冯耀煌, 杨旭. 论最优预报因子与最优预报方程. 气象学报, 1989, 47(1): 52—60.
- [5] 黄嘉佑. 我国月降水频数的时空特征. 气象, 1987, 13(1): 10—14.
- [6] 黄嘉佑. 转动主分量分析在天气气候分析中的应用. 气象, 1988, 14(9): 47—51.

THE EXPERIMENTS FOR SELECTING PREDICTORS FOR DAILY PRECIPITATION FORECASTING WITH MOS DURING FLOOD PERIOD IN SAN-HUA AREA OF YELLOW RIVER

Huang Jiayou

(Department of Geophysics, Peking University, Beijing 100871)

Fu Changfeng

(Henan Research Institute of Meteorological Science, Zhengzhou 450003)

· Abstract

In order to make daily forecasting for the precipitation with MOS in the San-Hua area of Yellow River in the flood period of 15 July to 15 August, the experiments for selecting the predictors from a lot of variables of model output statistics of numerical forecast proposed, which are four models: stepwise regression of linear correlation, nonlinear stepwise regression, nonlinear stepwise regression with principal component and rotated principal component. The results show that the model of nonlinear stepwise regression with rotated principal component has the highest skill score of forecast in the models and is synoptic meaningful.