

用车贝雪夫多项式做时间序列预报的 一种改进方案*

周家斌

(中国科学院大气物理研究所, 北京 100080)

提 要

为提高预报准确率, 本文对作者以前提出的时间序列预报方法, 提出了一种改进方案。这一方案的要点是: 1) 在计算理想初值的预报时, 针对不同情况选择不同的迭代次数; 2) 在通过迭代终值计算预报时, 对不同情况使用不同的拟合误差平均值; 3) 当拟合误差的符号难以确定时, 直接取迭代终值作为预报值。实例计算表明, 新的方案有利于提高预报准确率。

关键词: 时间序列; 预报方法

一、引 言

在 1982 年长江流域长期水文气象预报讨论会上, 作者提出了一种新的时间序列预报方法。此后, 我们将这一方法用于北京 4—8 月逐月降水量预报, 并将预报提供给北京市气象台。以后又应新安江电厂的要求向该厂提供了新安江流域年降水趋势预报。在实践过程中, 我们又对原来的预报方案^[1]做了进一步的完善^[2]。

八九年来的预报实践证明, 这一方法有一定的预报准确度。作为例子, 图 1 给出关于新安江流域的预报检验结果。由图可见, 1983—1990 年 8 年中, 除 1986 年外, 其余 7 年预报的降水趋势均正确。据新安江电厂生产技术科按水电部门评分标准评定¹⁾, 除 1986 年外, 其余 7 年均评为满分。北京市气象台对我们 1982—1989 年提供的预报进行了评定²⁾, 结果认为: 这一方法的预报效果逐步趋于稳定。1987 年以来预报准确率高于业务预报中所用的单一工具的预报准确率, 已做为一个有效预报工具投入日常业务使用。自这一方法提出以来, 已有许多气象水文台站将其用于业务预报。

尽管本方法有一定的准确率, 但在预报实践中也发现需要进一步改进。本文的目的就在于提出一种改进方案。为叙述和比较方便, 在第二节, 我们首先结合一个例子简述原来的计算方案, 第三节就是新提出的改进方案, 第四节是该方案的操作步骤。

1990 年 11 月 22 日收到, 1992 年 1 月 29 日收到修改稿

* 国家自然科学基金项目

1) 新安江电厂生产技术科给大气物理所的信 (1990)。

2) 北京市气象台对周家斌提供的预报的评定 (1990)。

二、原 方 案

这一预报方法的基本预报公式是^[1]

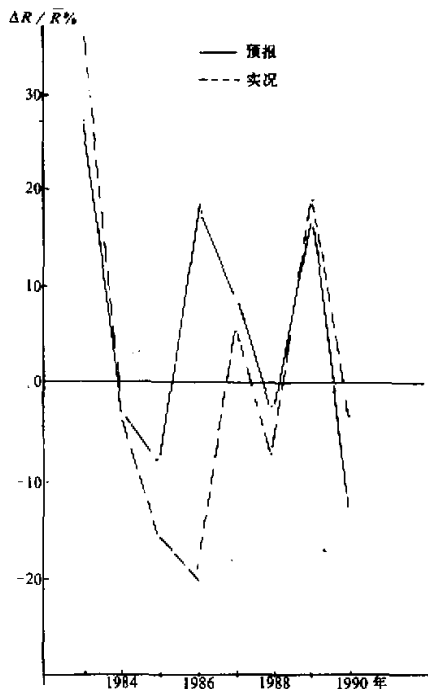
$$Z^{(v)}(T_0 + 1) = \sum_{k=0}^{K_0} A_k^{(v)} \varphi_k(T_0 + 1), \quad (1)$$

$$A_k^{(v)} = \sum_{t=1}^{T_0} Z(t) \varphi_k(t) + Z^{(v-1)}(T_0 + 1) \varphi_k(T_0 + 1), \quad (2)$$

式中 Z 为预报量, t 为时间标志。(1), (2)两式为一组迭代公式, v 为迭代次数。此二式用于由 $t=1, 2, \dots, T_0$ 时刻的资料 $Z(1), Z(2), \dots, Z(T_0)$, 计算 T_0+1 时刻的 $Z(T_0+1)$ 预报值, $\varphi_k(t)$ 为 k 阶归一化车贝雪夫多项式在格点 t 的值, A_k 为车贝雪夫系数, $A_k^{(v)}$ 为其 v 次迭代值。给定初值 $Z^{(0)}(T_0+1)$ 之后, 可由(1)、(2)两式依次循环求出各次近似值 $Z^{(1)}(T_0+1), Z^{(2)}(T_0+1), \dots, Z^{(v)}(T_0+1)$ 。上述迭代公式是收敛的^[1]。其迭代终值为

$$Z^{(\infty)}(T_0 + 1) = Z(T_0 + 1) + \frac{\hat{\varepsilon}}{1 - \sigma}, \quad (3)$$

图1 新安江流域年降水趋势预报检验
纵坐标为降水距平百分率



其中

$$\hat{\varepsilon} = \hat{Z}(T_0 + 1) - Z(T_0 + 1), \quad (4)$$

$$\hat{Z}(T_0 + 1) = \sum_{k=0}^{K_0} A_k \varphi_k(T_0 + 1), \quad (5)$$

$$A_k = \sum_{t=1}^{T_0+1} Z(t) \varphi_k(t), \quad (6)$$

$$\sigma = \sum_{k=0}^{K_0} \varphi_k^2(T_0 + 1). \quad (7)$$

式中 $\hat{Z}(T_0+1)$ 为当 $Z(1), Z(2), \dots, Z(T_0), Z(T_0+1)$ 皆已知时车贝雪夫展开式的拟合值, $\hat{\varepsilon}$ 为其拟合误差。

图2给出迭代终值 $Z^{(\infty)}(T_0+1)$ 与真值 $Z(T_0+1)$ 的关系。图中的 $\hat{\varepsilon} > 0$, $\hat{\varepsilon} < 0$ 的情况类之。此图表示 $Z^{(\infty)}(T_0+1), Z(T_0+1)$ 等在数轴上的位置, 左方以 $Z^{(\infty)}(T_0+1)$ 等为标度, 右方以 $\varepsilon^{(\infty)} (= Z^{(\infty)}(T_0+1) - Z(T_0+1))$ 等为标度, 以说明有关量之间的关系。

下面就是原来的计算方案。这一方案包括两个部分。

1. 由理想初值计算预报值

满足以下条件的初值为理想初值:

1) 初值误差 $\varepsilon^{(0)}$ 与拟合误差 $\hat{\varepsilon}$ 反号;

2) $|\varepsilon^{(0)}| > \frac{|\hat{\varepsilon}|}{\sigma}$. (8)

理想初值与 $\hat{\varepsilon}$ 及 $Z(T_0 + 1)$ 的关系见图 2。由此看来, 只要 $\varepsilon^{(0)}$ 与 $\hat{\varepsilon}$ 反号, 且 $|\varepsilon^{(0)}|$ 足够大, 则其相应初值就是理想的。

在实际计算中, 我们采用如下标准化值进行计算:

$$Z = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \times 100, \quad (9)$$

式中 Y 为未标准化的值, $Y_{\max}$, $Y_{\min}$ 为其最大最小值, Z 为标准化值。标准化后, $0 \leq Z \leq 100$ 。对标准化值, 若 $\hat{\varepsilon} \geq 0$, 取 $Z_i^{(0)}(T_0 + 1) = -400$, 若 $\hat{\varepsilon} < 0$, 取 $Z_i^{(0)}(T_0 + 1) = 500$ 。下标 i 表示初值是理想的。

应用理想初值进行迭代运算时, 经少数几次迭代, $Z^{(v)}(T_0 + 1)$ 即接近真值。接近真值时的迭代次数称临界迭代次数 v_c 。若 $\hat{\varepsilon} > 0$, 则迭代误差 $\varepsilon^{(v)} = Z^{(v)}(T_0 + 1) - Z(T_0 + 1)$ 在临界迭代次数附近由负号变为正号。若 $\hat{\varepsilon} < 0$, 则 $\varepsilon^{(v)}$ 在临界迭代次数附近由正号变为负号。图 2 中给出用理想初值迭代时预报值及其误差的变化情况。

每个个例的 v_c 不同, 为了实际预报的需要, 采用所有个例的最常见临界迭代次数为实用迭代次数, 以 v_c^* 表示。

对历史资料, 可由(4) - (6)式求 $\hat{\varepsilon}$ 。但在实际做预报时, 由于 $Z(T_0 + 1)$ 未知, 因而 $\hat{\varepsilon}$ 亦是未知的。为解决此困难, 可做出迭代终值分布图, 用该图来判断 $\hat{\varepsilon}$ 的符号。

图 3 给出一个例子。中央气象台把全国划分为 15 个区, 对每个区分别确定兼顾降水量和降水分布的降水指数。本例是第四区指数的计算结果。其中 $T_0 + 1 = 7$, $K_0 = 1$ 。图 3 中圆点为历史资料中 $\hat{\varepsilon} \geq 0$ 的个例, 三角形为 $\hat{\varepsilon} < 0$ 的个例。如图所示, 除少数个例外, 在临界值 $Z_c^{(\infty)}(T_0 + 1)$ 线以上, $\hat{\varepsilon} \geq 0$, 在其下, $\hat{\varepsilon} < 0$ 。在实际预报时, 即用此图确定 $\hat{\varepsilon}$ 的符号。

表 1 以 $Z^{(\infty)}(T_0 + 1)$ 的大小为序给出理想初值的计算结果。计算时, 在 $Z_c^{(\infty)}(T_0 + 1)$ 线以上的个例, 一律取 $Z_i^{(0)}(T_0 + 1) = -400$, 在其下, 取 $Z_i^{(0)}(T_0 + 1) = 500$ 。所有个例一律取 $v_c^* = 4$ 。 $Z_c^{(\infty)}(T_0 + 1)$ 以上的三角形和其下的圆点, 属于错判个例。对错判个例, 其相应初值将是非理想初值^[3]。对于这种非理想初值, 其迭代计算结果误差很大。为方便计, 对错判个例另行集中排序 (见表 1 第 25—33 行)。

具体计算结果见表 1 第 7 列, 相应的误差见第 11 列。

2. 从迭代终值出发做预报

预报公式如下:

$\hat{\varepsilon} \geq 0$ 时,

表 1 预报计算结果 (第四区降水, $T_0+1=7, K_0=1$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	区 间		年	Z	$\hat{\varepsilon}$	$Z^{(20)}$	Z_1		Z_α		ε_1		ε_α	
							旧	新	旧	新	旧	新	旧	新
1	A	A1	1965	30.8	37.6	101.0	77.7	50.9	76.3	48.7	46.9	20.1	45.5	17.9
2			1957	49.4	33.4	93.0	70.2	43.8	68.4	40.8	20.8	5.6	19.0	-8.6
3			1974	46.2	22.9	89.0	66.3	40.1	64.3	36.7	20.1	-6.1	18.1	-9.5
4		A2	1958	57.1	11.1	77.8	55.6	55.6	53.1	54.1	-1.5	-1.5	-4.0	-3.0
5			1966	49.4	12.7	73.0	51.1	51.1	48.3	49.3	1.7	1.7	-1.1	-0.1
6			1978	67.3	2.3	71.5	49.6	49.6	46.8	47.8	-17.7	-17.7	-20.5	-19.5
7			1972	11.5	31.1	69.6	47.8	47.8	44.9	45.9	36.3	36.3	33.4	34.4
8			1975	52.6	6.3	64.4	42.8	42.8	39.7	40.7	-9.8	-9.8	-12.9	-11.9
9	B	B+	1960	51.9	3.0	57.6	36.3	57.6	32.9	57.6	-15.6	5.7	-19.0	5.7
10			1967	54.5	1.7	57.6	36.3	57.6	32.9	57.6	-18.2	3.1	-21.6	3.1
11			1979	35.9	11.4	57.2	35.9	57.2	32.5	57.2	0.0	21.3	-3.4	21.3
12			1961	39.1	9.2	56.2	35.0	56.2	31.5	56.2	-4.1	17.1	-7.6	17.1
13			1980	14.1	20.7	52.8	31.7	52.8	28.1	52.8	17.6	38.7	14.0	38.7
14			1986	14.1	18.2	48.2	27.3	48.2	23.5	48.2	13.2	34.1	9.4	34.1
15			1985	35.3	5.9	46.3	25.5	46.3	21.6	46.3	-9.8	11.0	-13.7	11.0
16			1968	0.0	21.0	39.2	18.9	39.2	14.5	39.2	18.9	39.2	14.5	39.2
17		B-	1971	92.3	-30.6	35.1	56.7	35.1	64.7	35.1	-35.6	-57.2	-27.6	-57.2
18			1962	69.2	-19.2	33.4	55.1	33.4	63.0	33.4	-14.1	-35.8	-6.1	-35.8
19	C	C2	1981	33.3	-6.5	21.2	43.4	43.4	50.8	49.0	10.1	10.1	17.5	15.7
20			1988	44.9	-15.9	15.3	37.8	37.8	44.9	43.1	-7.1	-7.1	0.0	-1.8
21			1984	57.1	-22.4	15.2	37.7	37.7	44.8	43.0	-19.4	-19.4	-12.3	-14.1
22		C1	1982	48.7	-19.6	12.2	34.0	61.0	41.8	51.8	-13.9	12.3	-6.9	3.1
23			1970	48.7	-24.5	3.0	26.1	52.8	32.6	42.6	-22.6	4.1	-16.1	-6.1
24			1969	34.0	-19.5	-2.4	21.0	38.1	27.2	37.2	-13.0	4.1	-6.8	3.2
25	A	A2	1964	100.0	-11.8	78.0	55.8	55.8	53.3	54.3	-44.2	-44.2	-46.7	-45.7
26	B	B+	1963	89.1	-16.4	58.5	37.2	58.5	33.8	58.5	-51.9	-30.6	-55.3	-30.6
27			1977	57.1	-0.2	56.7	35.5	56.7	32.0	56.7	-21.6	-0.3	-25.1	-0.3
28			1959	64.1	-5.0	54.8	33.7	54.0	30.1	54.8	-30.4	-9.3	-34.0	-9.3
29			1976	69.2	-8.7	52.9	31.9	52.9	28.2	52.9	-37.3	-16.3	-41.0	-16.3
30			1973	90.4	-22.8	47.9	27.1	47.9	23.2	47.9	-63.3	-42.5	-67.2	-42.5
31		B-	1989	21.2	6.4	33.1	54.8	33.1	62.7	33.1	36.6	11.9	41.5	11.9
32			1983	14.7	6.1	26.2	48.2	26.2	55.8	26.2	33.5	11.5	41.1	11.5
33			1987	20.5	2.1	24.4	46.5	24.4	54.0	24.4	26.0	3.9	33.5	3.9

A1, A2, C1, C2 为区间 A, C 中的子区间, B+和 B-分别为 B 区间中 $\hat{\varepsilon} \geq 0$ 和 $\hat{\varepsilon} < 0$ 的个例子区间, "旧"表示原方案计算结果, "新"表示改进方案计算结果。

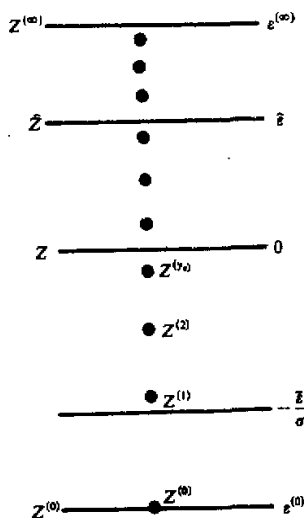


图2·真值 Z 、拟合值 $\hat{Z}$ 、拟合误差 $\hat{e}$ 、迭代终值 $Z^{(\infty)}$ 、理想初值 $Z^{(0)}$ 、理想初值误差 $e^{(0)}$ 的关系及用理想初值进行计算时的迭代结果($Z^{(1)}, Z^{(2)}, \dots, Z^{(v_0)}, \dots$)的变化($\hat{e} > 0$)

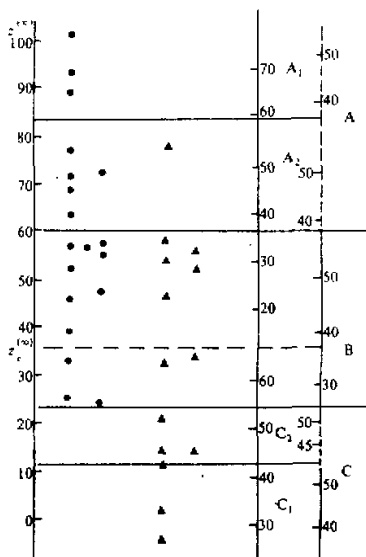


图3 迭代终值分布及预报工具图

(第四区降水, $T_0 + 1 = 7, K_0 = 1$)

(此图表示各个例迭代终值在数轴上的分布, 为避免点子过分密集, 将数值相近的点做了平移)

$$Z_{\infty}(T_0 + 1) = Z^{(\infty)}(T_0 + 1) - \frac{\bar{\hat{e}}_+}{1 - \sigma}, \quad (10)$$

$\hat{e} < 0$ 时,

$$Z_{\infty}(T_0 + 1) = Z^{(\infty)}(T_0 + 1) - \frac{\bar{\hat{e}}_-}{1 - \sigma}, \quad (11)$$

式中 $Z_{\infty}(T_0 + 1)$ 为预报值, $\bar{\hat{e}}_+$ 和 $\bar{\hat{e}}_-$ 分别为 $\hat{e} \geq 0$ 和 $\hat{e} < 0$ 时的平均值。对图3中 $Z_c^{(\infty)}$ 线以上的个例, 用(10)式计算, 对线下的个例, 用(11)式计算。结果见表1第9列, 相应的误差见第13列。显然, 错判个例误差很大。

三、改进方案

分析上节计算过程及其结果, 可以看到, 预报误差的主要来源是: 1) 对 $\hat{e}$ 的符号判错。2) 在用理想初值计算时, $\hat{e}$ 的符号判断虽然正确, 但由于适合个例的临界迭代次数 v_c 与预报计算用的统一迭代次数 v_c^* 不同, 将造成较大的预报误差。3) 在通过迭代终值计算预报时, 即使 $\hat{e}$ 的符号判断正确, 但由于个例的 $\hat{e}$ 值与公式(10)和(11)中使用的平均值不同, 将造成预报误差(当 $\hat{e}$ 与平均值相同时, 由(3)、(10)、(11)式得知, 预报将不存在误差)。改进方案就是从减少以上误差的目的提出的。

首先将迭代终值分布图分成不同的区间,采用不同的处理方法。例如图 3,分成三个区间。在区间 A 和区间 C,我们可以正确(在区间 C)或基本上正确(在区间 A)判断 $\hat{\varepsilon}$ 的符号。在区间 B,则有一部分个例判错。现分区间叙述处理方法。

1. 区间 B

在这个区间, $\hat{\varepsilon} \geq 0$ 和 $\hat{\varepsilon} < 0$ 的个例混在一起,不易正确判断 $\hat{\varepsilon}$ 的符号。为了避免因误判引起的大预报误差,在这个区间,我们直接取 $Z^{(\infty)}(T_0 + 1)$ 为预报值。

(3) 式中 $Z^{(\infty)}(T_0 + 1)$ 由 $Z(T_0 + 1)$ 和 $\hat{\varepsilon}$ 两个因子决定。一般说来,该区间 $\hat{\varepsilon}$ 较小(参看表 1 第 5 列)因而对 $Z^{(\infty)}(T_0 + 1)$ 的影响小, $Z(T_0 + 1)$ 与 $Z^{(\infty)}(T_0 + 1)$ 的值比较接近。因此,上述处理是可行的。由此做出的预报见表 1 第 8 列、10 列第 9—18 和 26—33 行,相应的误差见第 12 列、14 列相应行。

2. 区间 A 和区间 C

在此二区间中, $Z(T_0 + 1)$ 的值差别不大, $Z^{(\infty)}(T_0 + 1)$ 的值基本上由 $\hat{\varepsilon}$ 决定。考虑到这一点,我们再把区间 A 和区间 C 各自划分为若干子区间(如子区间 A1, A2, C1, C2)处理。

1) 用理想初值做预报

分析每个个例的临界迭代次数,发现离 $Z_c^{(\infty)}(T_0 + 1)$ 距离愈远,临界迭代次数愈少。为此,在不同的子区间中采用不同的迭代次数。各子区间的迭代次数见表 2。为了更清楚地表示这一点,我们在表 2 中还给出了另一组计算结果(对本文例, $K_0 = 1$,另一例 $K_0 = 2$)。表 1 第 8 列 1—8 行和 19—24 行就是计算结果,第 12 列的相应行为预报误差。

2) 通过迭代终值做预报

在不同的子区间中,分别计算 $\hat{\varepsilon}$ 的平均值,并用以下公式计算预报:

$\hat{\varepsilon} \geq 0$ 时

$$Z_{\infty}(T_0 + 1) = Z^{(\infty)}(T_0 + 1) - \frac{\bar{\hat{\varepsilon}} + k}{1 - \sigma}, \quad (12)$$

$\hat{\varepsilon} < 0$ 时

$$Z_{\infty}(T_0 + 1) = Z^{(\infty)}(T_0 + 1) - \frac{\bar{\hat{\varepsilon}} - k}{1 - \sigma}, \quad (13)$$

式中 $\bar{\hat{\varepsilon}}$ 的平均值有一下标 k ,表示不同子区间的平均值不同。这些平均值 $\bar{\hat{\varepsilon}}$ 见表 2。在区间 A,用(12)式,在区间 C 用(13)式。相应的计算结果见表 1 第 10 列第 1—8 行和 19—24 行,第 14 列相应行为预报误差。

我们对新旧方案的预报误差进行了总体比较,表 3 就是预报误差的比较结果。可以看到,新方案的预报误差比原方案要小。

图 4 给出 $K_0 = 1$ 时由迭代终值出发做预报时两种方案计算结果的比较。从图中很清楚地看到新方案的计算结果(虚线)比原方案的(锁线)更接近实况(实线)。在 35 个个例中,新方案对绝大多数个例有改进,其中尤以 1959、1960、1967、1976、1977、1983、1987、1989 诸年为最显著。

表 2 不同子区间的迭代次数 v_e 和 $\bar{\varepsilon}$ 的平均值 (第四区, $T_0 + 1 = 7$)

K_0	子区间	v_e	$\bar{\varepsilon}$
1	A1	3	28.0
	A2	4	12.7
	C2	4	-14.9
	C1	3	-21.2
2	A1	8	17.2
	A2	10	6.6
	C2	10	-7.4
	C1	8	-13.0

表 3 新旧方案误差比较 (第四区, $T_0 + 1 = 7$)

K_0	$ \varepsilon_1 _{旧}$	$ \varepsilon_1 _{新}$	$ \varepsilon_\infty _{旧}$	$ \varepsilon_\infty _{新}$
1	22.1	17.9	22.3	17.7
2	23.2	20.2	24.1	21.16

$|\varepsilon_1|$ 为用理想初值做预报时绝对误差的平均值, $|\varepsilon_\infty|$ 为从迭代终值出发做预报时的相应误差。新方案中 B 区间皆取迭代终值为预报值, “新”, “旧”意义同表 1。

这里, 我们顺便指出, 本文的改进方案也适用于资料的插补^[4]。

四、改进方案的操作步骤

现在, 我们把改进方案的操作步骤整理如下:

1. 历史资料的处理

1) 资料预处理: 由原始资料求出标准化值。

2) 确定格点数 $T_0 + 1$ 。

3) 拟合, 求 $\hat{\varepsilon}$ 。

4) 由 (3) 式计算迭代终值。

5) 制作迭代终值分布图。依照点的分布情况将图分成 A, B, C 三个区间, 在区间 A, C 中再依 $\hat{\varepsilon}$ 的分布划分若干子区间。

6) 在区间 B 中取迭代终值为预报值。

7) 在区间 A, C 中用理想初值进行迭代, 分子区间确定迭代次数并做出理想初值的预报。

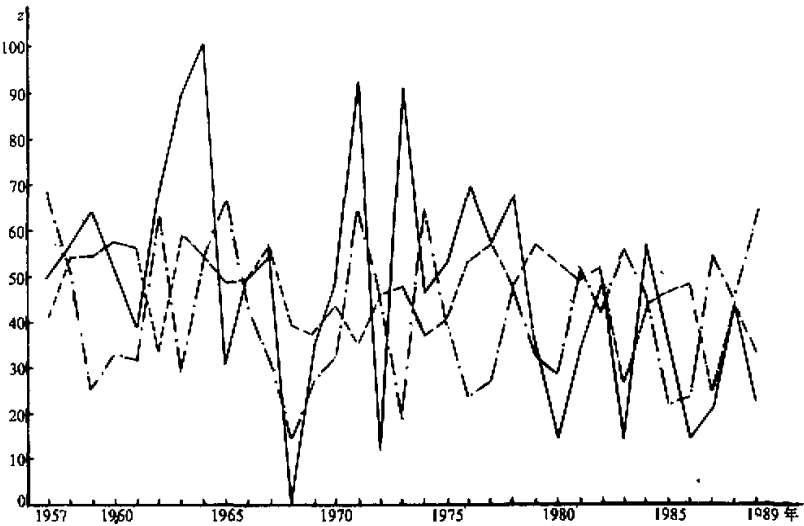


图 4 新旧方案预报结果比较 (第四区降水, $T_0 + 1 = 7$, $K_0 = 1$)
——新方案预报, - - -旧方案预报,实况。

8) 在区间 A、C 的各个子区间中分别确定 $\hat{\varepsilon}$ 的平均值,并用 (12)、(13) 式做出相应的预报。

9) 计算预报误差并进行统计。

2. 实际预报

1) 经多次迭代在一定的误差控制下求出迭代终值,并将结果点入由历史资料绘出的迭代终值分布图中。

2) 若计算所得的迭代终值位于 B 区间中,此值即为预报值。

3) 若迭代终值不在区间 B,则取理想初值做预报。迭代次数以点所在的子区间历史资料个例的迭代次数为准。

4) 若点在区间 A,用 (12) 式计算预报,若在区间 C,用 (13) 式计算预报。平均值用点所在子区间中历史资料个例 $\hat{\varepsilon}$ 的平均值。

参 考 文 献

- [1] 周家斌, 1983, 一种新的时间序列预报方法, 科学通报, 28, 1250-1252.
- [2] 周家斌, 1985, 关于用车贝雪夫多项式做时间序列预报的几个问题, 科学通报, 30, 1163-1166.
- [3] 周家斌, 1990, 车贝雪夫多项式及其在气象中的应用, 第九章, 气象出版社.
- [4] 周家斌, 1987, 一种气象资料插补方法, 科学通报, 32, 1199-1200.

An Improved Scheme for the Forecast of Time Series by Using Chebyshev Polynomials

Zhou Jiabin

(Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029)

Abstract

To raise the accuracy of forecast an improved acheme is presented in this paper for the forecasting method of time series proposed by the author previously. The main points of the scheme are: (1) using different iteration numbers for different initial values of iteration; (2) using different means of fitting for predicted value obtained from final value of iteration; (3) taking the final value as prediction when it is difficult to difine the sign of final value of iteration. The results indicate that the accuracy of the forecast may be raised by using the improved scheme.

Key words: time series; forecasting method.