

# 人工神经网络的短期气候定性预测方法研究

覃志年<sup>1</sup> 金龙<sup>2</sup> 况雪源<sup>1</sup>

(1. 广西壮族自治区气候中心; 2. 广西壮族自治区气象减灾研究所, 南宁 530022)

**摘要** 采用 EOF 方法将众多具有一定物理意义的环境场预报因子和序列周期因子进行降维去噪处理, 并结合人工神经网络前馈网络模型, 进行了影响广西热带气旋年频数和夏季(6~8 月)降水量短期气候预测的定性预报建模方法研究。结果表明, 该方法对于影响广西热带气旋的年频数及广西夏季降水量定性趋势具有很好的预报能力, 可作为一种新的短期气候预测业务预报配套方法。

**关键词** 降维 神经网络 热带气旋频数 夏季降水 定性预报

## 引言

广西地处低纬, 西有高原, 南面临海, 天气气候复杂多变, 给本地区的短期气候业务预报工作带来了很大的难度。特别是诸如影响广西的热带气旋年频数、广西夏季(6~8 月)降水量等短期气候预测问题, 由于其本身数值的年际变化很大以及影响这类长期天气过程的机理十分复杂, 因此, 其预报难度更大。为了探索新的有效预报方法, 本文尝试在合理选取环境场及序列周期因子的情况下, 应用自然正交函数(EOF)展开方法, 将通过一定信度检验初选出来的众多预报因子“浓缩”成少数几个新的预报因子变量, 再结合人工神经网络方法, 对影响广西的热带气旋年频数及广西夏季降水趋势, 进行短期气候预测方法研究。

## 1 预报因子的选取与分析处理

预报因子的选取是短期气候统计预报的重要基础。在预报方法相同的情况下, 因子物理意义清楚, 独立性强, 其预报结果稳定可靠。在气象学科中, 很多气象要素的变化, 除了受到海洋与大气的相互作用外, 还取决于其自身活动规律。为此, 在预报因子的组合上, 有必要充分考虑环境场与预报量自身周期因子。

### 1.1 环境场因子分析

本文在选取环境场因子时, 基本因子场包括两大类: 一是各种月平均环流指数(包括副高指数、欧亚环流指数及从网上下载的各种指数、振幅位相、环流型天数等), 二是北太平洋月平均海温场及北半球 500 hPa 和 100 hPa 高度场(以下简称“三场”)。考虑到各种资料起始年代不一及预报发布的时间(每年 3 月底之前), 前期因子一律从上年 1~12 月资料查找(以下称为上下年关系因子), 资料起始时间均从 1956 年开始。而在选取因子时, 同种资料每月仅取达到信度检验标准以上的相关系数最大的一个因子。

在应用第二类因子计算相关时, 先对三场分别作如下预处理: 如使用每月海温场的 286 个格点作相关计算时, 为了使其与实际地理位置相对应及处理方便, 将其变为  $13 \times 33$  个格点矩阵(不足部分用“0”填充)。然后从北向南、从西向东, 以  $4 \times 4$  个格点范围进行滑动块状求其平均值(矩阵中的“0”是不参与计算的)。而对于 500 hPa 及 100 hPa 高度场, 依照上述顺序分别以  $4 \times 5$  和以  $4 \times 4$  格点范围求其滑动块状的平均值。经过这样处理可得到具有一定大小范围的新的三场。而每一个新的值, 容易和天气形势或天气系统建立一定的联系。因此, 找出的因子具有较好的代表性和稳定性。

国家自然科学基金项目(40075021)资助

作者简介: 覃志年, 男, 1955 生, 高级工程师, 从事短期气候预测工作

收稿日期: 2003 年 6 月 5 日; 定稿日期: 2003 年 10 月 26 日

现以影响广西热带气旋年频数预报为例,通过 0.05 以上的因子(表 1)。实际预报因子普查,共得到 38 个信度大于等于

表 1 影响广西热带气旋年频数的前期因子各月最大相关系数

|     | 1 月   | 2 月   | 3 月 | 4 月   | 5 月   | 6 月   | 7 月   | 8 月   | 9 月   | 10 月  | 11 月  | 12 月  |
|-----|-------|-------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| H50 | -0.42 | -0.36 |     | -0.37 | -0.52 | -0.60 | -0.46 | -0.46 | -0.60 | -0.38 | -0.38 | -0.45 |
| SST | -0.39 | 0.38  |     | 0.35  | 0.36  | 0.37  | 0.40  |       |       | -0.36 |       |       |
| H10 | -0.38 | -0.35 |     |       | -0.35 | -0.38 | -0.48 | -0.40 | -0.55 | -0.50 | -0.57 | -0.40 |
| ZF  |       | -0.35 |     |       | -0.35 |       |       | -0.38 |       |       |       | -0.40 |
| WX  | -0.38 |       |     |       | -0.41 |       | 0.36  | 0.37  | 0.40  |       |       |       |
| WPP |       |       |     |       |       |       |       |       | 0.36  |       |       |       |

注:H50、SST、H10 分别为 500 hPa 海温、100 hPa 块状高度场,ZF、WX为振幅位相,WPP为 West Pacific Pattern 环流型缩写

由表 1 可见,多数的因子都在“三场”之中,ZF、WX 等其它指数很少有通过 0.05 信度相关的因子。而在“三场”的因子中,500 hPa 的 5、6、9 月及 100 hPa 的 9、10、11 月的因子相关显著水平高达 0.01。从 500 hPa 相关的地理位置看,上年春末到盛夏位于东西伯利亚地区有一较稳定的高相关区,而到上年秋季,在格陵兰西部有一稳定且明显的反相关区。说明上年春末到盛夏,若东亚阻高不明显及秋季格陵兰西部有一稳定的强低涡,来年影响广西热带气旋年频数则偏多。图 1 给出的是上年 9 月 500 hPa 格陵兰西部高度距平场与影响广西热带气旋年频数 (R = -0.6) 各自排序后的相关图。可见,格陵兰西部高度距平场与广西热带气旋年频数的相关较为显著。

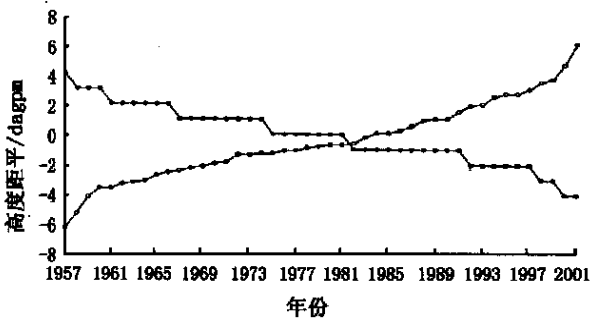


图 1 格陵兰西部高度距平场(实心圆)与影响广西的热带气旋年频数(空心圆)各自排序后的相关图

1.2 周期因子

由于受天文等因素的强迫作用,很多气象要素存在着不同时间尺度的周期变化,天气过程也有再现的情况。因此,用周期制作预报有时会得到较好的效果。但是,在制作某气候要素的预报中,如果仅考虑单方面环境场因子或自身周期因子,其预报方程或模式所包含的信息很可能不够全面,预报准确率也容易受到限制和不稳定。为此,本文在建立预

报模型时,在利用环境场因子的基础上增加了周期因子。周期因子是利用均生函数计算方法<sup>[1]</sup>,即对预报量序列计算均生函数:

$$\bar{x}_k(i) = \frac{1}{n_k} \sum_{j=0}^{n_k-1} x(i + jk) \quad (1)$$

式中  $i = 1, 2, \dots, k, 1 \leq k \leq M$  且  $n_k = \text{int}(N/k)$ ,  $M$  取  $\text{int}(N/2)$ 。进一步对均生函数作周期性延拓计算:

$$f_k(i) = \bar{x}_k \left| i - k \text{int} \left\lfloor \frac{i-1}{k} \right\rfloor \right| \quad (2)$$

其中  $i = 1, 2, \dots, N, k = 1, 2, \dots, M$ 。

将均生函数延拓序列与预报量计算相关,并选取一个长周期一个短周期作为周期因子。其中,在周期为 2 ~ 9 之间选取相关系数最大的一个作为短周期因子(本例短周期为 8 年),在周期为 10 ~ N/2 之间选取相关系数最大的一个作为长周期因子(本例长周期为 23 年)。

1.3 因子降维处理

在多元分析的各种实际预报方法中,都存在着一个共性的问题,即在建模预测时,如何从大量的初选因子中精选出最终用于建立预报方程的预报因子。目前通常采用的几种方法是:①逐个引进因子的向前筛选法,该方法每次将残差平方和减小最多的因子引入方程,而且方程中的因子数一般不超过 10 个;②采用因子逐个淘汰的后向筛选法,即将所有因子建立方程,然后把对预报量贡献小的因子逐个剔除;③因子有进有出的逐个筛选法,即按一定标准逐个引进因子的同时,把因新因子的引入而不符合标准的因子逐个剔除。

显而易见,采用上述 3 种方法,均会使初选的大量预报因子中的很多因子,无法进入最终的预报方程,从而丢失和浪费了很多有用的预报信息,也无法

更全面地反映预报因子与预报量的关系;另一方面,如果采用较多的预报因子建立神经网络的预报模型时,面对庞大的因子群,不仅计算速度慢得难以忍受,同时也会增加噪声和计算误差。为了弥补这一缺陷,本文利用自然正交函数展开(EOF)的方法<sup>[2]</sup>,将通过一定标准选出的大量因子降至少数几个新的因子变量,用来建立预报模型。即设有  $m$  个初选因子:

$${}_n X_m = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad (3)$$

经过线性变换可以分解成只依赖于时间与空间的函数:

$${}_n X_m = {}_n Z_m \Phi_m \quad (4)$$

其中,  $\Phi_m$  为空间函数,  ${}_n Z_m$  为时间函数。由于 EOF 收敛速度快、浓缩性强,并且所分解出来的特征向量是正交的。因此,一般用前面少数几个时间函数,基本可以代表原来因子群的主要信息。这样既最大限度地利用了全部预报因子的有用预报信息,也避免了由于因子过多导致误差积累造成方程式的退化。

## 2 神经网络定性预报的主要算法

一种好的预报方法往往可以更好地解释和利用预报因子。人工神经网络方法,是近年来新兴的一种非线性预报方法。与传统的线性顺序处理技术相比,神经网络方法更适合应用于具有非线性演变特征的旱涝等气象灾害的长期预报问题。作者曾采用神经网络方法进行了月降水量、农田土壤水分等预报建模试验,而以往在大气学科中应用最广泛的是采用神经网络的单输出模型进行神经网络的定量预报<sup>[3~5]</sup>。本文则尝试应用神经网络的多输出模型进行短期气候的定性预报研究,该方法的主要数学原理及思路为:

设网络的学习样本输入和期望输出  $A_k, C_k$  ( $k = 1, 2, \dots, m$ ) 为学习矩阵的模式输入对个数。随机地给出网络从输入层到隐含层的初始连接权系数  $V_{hi}$ ,同时也随机地给出一组隐含层到输出层的连接权系数  $W_{ij}$  以及隐含层单元和输出层单元的阈值  $\theta_i, \gamma_j$ 。然后对学习样本的输入和期望输出  $A_k, C_k$  ( $k = 1, 2, \dots, m$ ) 进行以下计算。

① 根据连接权矩阵和学习样本的输入,计算隐含层的激活值:

$$b_i = f\left(\sum_{h=1}^n a_h V_{hi} + \theta_i\right) \quad (5)$$

其中  $i = 1, 2, \dots, p$  为隐含层节点数,  $h = 1, 2, \dots, n$  为输入层节点数。节点数的转移函数为 Sigmoid 函数:

$$f(x) = 1/(1 + e^{-x}) \quad (6)$$

② 计算输出层单元的激活值:

$$C_j = f\left(\sum_{i=1}^p W_{ij} b_i + \gamma_j\right) \quad (7)$$

其中  $j = 1, 2, \dots, q$  为输出层节点数,初始时刻  $W_{ij}$  为一组给定的随机变量。

③ 计算输出层单元的一般化误差

$$d_j = C_j(1 - C_j)(C_j^k - C_j) \quad (8)$$

其中  $j = 1, 2, \dots, q$ ,  $C_j^k$  为输出层单元  $j$  的期望输出。

④ 计算隐含层单元相对每个  $d_j$  的误差:

$$e_i = b_i(1 - b_i) \sum_{j=1}^q W_{ij} d_j \quad (9)$$

其中  $i = 1, 2, \dots, p$ 。

⑤ 调整隐含层单元到输出层单元的连接权:

$$\Delta W_{ij} = \alpha b_i d_j \quad (10)$$

其中  $i = 1, 2, \dots, p$  和  $j = 1, 2, \dots, q$ ,  $\alpha$  为学习因子 ( $0 < \alpha < 1$ )。

⑥ 调整输出层单元的阈值:

$$\Delta \gamma_j = \alpha d_j \quad (11)$$

其中  $j = 1, 2, \dots, q$ 。

⑦ 调整输入层单元到隐含层单元的连接权:

$$\Delta V_{hi} = \beta \alpha_i e_i \quad (12)$$

其中  $h = 1, 2, \dots, n$ ,  $i = 1, 2, \dots, p$ ,  $\beta$  为动量因子 ( $0 < \beta < 1$ )

⑧ 调整隐含层单元的阈值:

$$\Delta \theta_i = \beta e_i \quad (13)$$

其中  $i = 1, 2, \dots, p$ 。

⑨ 重复第 ① 到第 ⑧ 的计算步骤,直到  $j = 1, 2, \dots, q$  和  $k = 1, 2, \dots, m$ , 计算实际输出与期望输出的误差。当全部样本的输出误差小于设定的收敛误差时,训练结束。根据这些确定的连接权系数和阈值,就可以得到神经网络的定性预报模型。

## 3 预报模型和检验分析

广西是受热带气旋影响最多的省区之一。据统

计,1951~2002年影响广西的热带气旋有290个之多,平均每年5.5个。在每年影响广西的热带气旋中,约有50%可以深入到广西内陆。计算表明,广西夏季雨量与影响广西的热带气旋频数关系密切( $R=0.55$ )。通过分析广西热带气旋典型偏多的年份及深入到广西内陆的热带气旋发现,其中95%以上影响广西的热带气旋都会产生暴雨、洪涝等灾害天气,带来的经济损失很大。因此,预报影响广西热带气旋的年频数极为重要。长期以来,当地预报工作者常把预报广西夏季雨量以及洪涝灾害的轻重年度与影响广西热带气旋年频数紧密联系,可见报准影响广西热带气旋年频数至关重要。然而,迄今为止,对于影响广西热带气旋的研究,大多是针对影响广西的热带气旋移动路径或其所造成的灾害天气等中短期预报方面,对于热带气旋年频数的预报方法研究较少。而广西夏季降水量占全年降水量的46%,同时也是洪涝灾害发生频率最高的时段。本文所建的预报模型,是以影响广西的热带气旋年频数和广西夏季雨量作为预报对象,同时也考虑到这两个预报量在实际业务工作中预报难度较大、预报准确率低,是短期气候业务预报的重点和难点问题。

在建立影响广西的热带气旋年频数的预测模型中,影响广西热带气旋年频数序列的预报建模资料样本为1957~1997年41个样本,并以1998~2002年5年资料作为对实际预测检验。

根据上述神经网络的学习算法,在建立预测模型时,采用一个3层的前馈网络模型,该模型包括一个输入层和一个输出层,其中间为一个隐含层。首先对初选出的40个因子(含2个周期因子)进行标准化处理及EOF展开,再将EOF展开后的40列时间函数与预报量计算相关,提取前3个最大相关的时间函数作为神经网络建模的学习矩阵输入。将学习矩阵加载到一个3层的前馈网络模型。其中,学习矩阵的输入节点数为3,输出节点数为2,即该学习矩阵的目标输出是将建模样本的预报量按正负距平分10(正距平),01(负距平),并且取隐节点数为3,动量因子为0.9,学习因子为0.7。对该学习矩阵作1000次训练,误差函数趋于稳定后,训练结束。表2给出了影响热带气旋频数神经网络定性预报模型对历史样本的趋势拟合结果。由表2可以看到,在41个历史样本中有4年拟合趋势错误,历史样本的拟合准确率达到90.2%。另外,在表2中还

给出了该预报模型对1998年到2002年5年的独立样本预报结果,从预报结果可以看到,5年中,预报对4年,错1年,预报准确率达到80%。

表2 影响广西的热带气旋拟合及预测结果

| 年份            | 实际输出   |        | 期望输出  |
|---------------|--------|--------|-------|
| 1957          | 0.0180 | 0.9820 | 0 1 ✓ |
| 1958          | 0.3051 | 0.6951 | 1 0 × |
| 1959          | 0.0719 | 0.9284 | 0 1 ✓ |
| 1960          | 0.9926 | 0.0073 | 1 0 ✓ |
| 1961          | 0.9847 | 0.0153 | 1 0 ✓ |
| 1962          | 0.0309 | 0.9692 | 0 1 ✓ |
| 1963          | 0.6128 | 0.3873 | 1 0 ✓ |
| 1964          | 0.9466 | 0.0532 | 1 0 ✓ |
| 1965          | 0.8670 | 0.1328 | 1 0 ✓ |
| 1966          | 0.0509 | 0.9493 | 0 1 ✓ |
| 1967          | 0.9925 | 0.0074 | 1 0 ✓ |
| 1968          | 0.4440 | 0.5563 | 0 1 ✓ |
| 1969          | 0.0202 | 0.9799 | 0 1 ✓ |
| 1970          | 0.6853 | 0.3145 | 0 1 × |
| 1971          | 0.9852 | 0.0148 | 1 0 ✓ |
| 1972          | 0.2512 | 0.7492 | 0 1 ✓ |
| 1973          | 0.9666 | 0.0333 | 1 0 ✓ |
| 1974          | 0.9861 | 0.0138 | 1 0 ✓ |
| 1975          | 0.9884 | 0.0116 | 1 0 ✓ |
| 1976          | 0.1076 | 0.8926 | 0 1 ✓ |
| 1977          | 0.0584 | 0.9417 | 0 1 ✓ |
| 1978          | 0.4816 | 0.5183 | 0 1 ✓ |
| 1979          | 0.5921 | 0.4077 | 0 1 × |
| 1980          | 0.2078 | 0.7923 | 0 1 ✓ |
| 1981          | 0.9419 | 0.0580 | 1 0 ✓ |
| 1982          | 0.0887 | 0.9116 | 0 1 ✓ |
| 1983          | 0.0807 | 0.9195 | 0 1 ✓ |
| 1984          | 0.0315 | 0.9686 | 0 1 ✓ |
| 1985          | 0.8859 | 0.1139 | 1 0 ✓ |
| 1986          | 0.1621 | 0.8382 | 1 0 × |
| 1987          | 0.0134 | 0.9866 | 0 1 ✓ |
| 1988          | 0.0410 | 0.9590 | 0 1 ✓ |
| 1989          | 0.5859 | 0.4141 | 1 0 ✓ |
| 1990          | 0.0111 | 0.9889 | 0 1 ✓ |
| 1991          | 0.0719 | 0.9282 | 0 1 ✓ |
| 1992          | 0.1329 | 0.8673 | 0 1 ✓ |
| 1993          | 0.0203 | 0.9797 | 0 1 ✓ |
| 1994          | 0.8343 | 0.1653 | 1 0 ✓ |
| 1995          | 0.6017 | 0.3981 | 1 0 ✓ |
| 1996          | 0.7193 | 0.2803 | 1 0 ✓ |
| 1997          | 0.0440 | 0.9563 | 0 1 ✓ |
| 平均拟合准确率:90.2% |        |        |       |
| 独立样本预报        |        |        |       |
| 1998          | 0.1069 | 0.8935 | 0 1 ✓ |
| 1999          | 0.0910 | 0.9092 | 0 1 ✓ |
| 2000          | 0.0345 | 0.9656 | 0 1 ✓ |
| 2001          | 0.0828 | 0.9175 | 0 1 ✓ |
| 2002          | 0.9746 | 0.0254 | 0 1 × |

依照上述同样的方法,进一步以广西夏季(6~8月)降水量为预报量,进行了预报建模试验。以信度 $\alpha=0.05$ 作为标准,共初选得到32个环境场预报因子。并同样根据均生函数计算方法选取1个短周期因子和1个长周期因子,共得到34个初选因子(预报因子表略)。对该34个因子作标准化处理和

EOF 展开,并计算 34 个时间函数与预报量的相关,从中提取 3 个相关最好的时间函数,作为神经网络建模的学习矩阵输入。对该学习矩阵作学习训练,同样学习因子取 0.7,动量因子取 0.9,隐节点数取 3,当学习矩阵作 1000 次训练后,误差函数趋于平稳,训练结束。结果表明,该预报模型对广西夏季降水 41 年历史样本的拟合准确率达到 87.8%,而对最近 5 年的独立样本预报,除 1999 年趋势预报错误,其余 4 年的预报结果均正确(表略)。

#### 4 小结

(1) 对于影响广西热带气旋年频数和广西夏季降水量两个预报对象的实际预测表明,本文提出的神经网络定性预报方法有较强的趋势预报能力,可作为短期气候定量预报的一种配套预报方法。

(2) 预报效果的好坏及稳定与否,与因子的选取及怎样有效利用好初选出来的因子群紧密相关。本文在选取具有一定物理意义的环境因子与周期因子相结合的情况下,通过 EOF 降维的方法,不仅充分利用了原来全部因子的主要有效预报信息,也起到

精选因子的作用。从而使人工神经网络定性预报模型计算速度大为加快,有利于人工神经网络定性预报方法投入业务使用。

(3) 在初选预报因子时,将海温及高空环流格点资料处理成块状平均场,所选出的因子容易与天气系统及环流形势建立联系,因而相关稳定,物理意义清楚;在此基础上,增加均生函数周期预报因子,更有利于提高预报准确率。

#### 参考文献

- 1 魏凤英,曹鸿兴. 长期预报的数学模型及其应用. 北京:气象出版社,1990.9-42
- 2 丁裕国. 经验正交函数展开气象场收敛性的研究. 热带气象, 1998,4(4):316-325
- 3 金龙,罗莹,缪启龙,等. 农田土壤湿度的人工神经网络预报模型研究. 土壤学报,1998,35(1):25-32
- 4 Jin Long,Jun Weimin,Miao Qilong. Study on AN-based multi-step prediction model of short-term climatic variation. Adv. Atmos. Sci.,2000,17(1):157-164
- 5 袁美英,潘华盛,张桂华. 利用人工神经网络预报松花江汛期水位. 气象科技,2002,30(1):43-46

## A Short-Term Climate Qualitative Forecast Model by ANN

Qin Zhinian<sup>1</sup> Jin Long<sup>2</sup> Kuang Xueyuan<sup>1</sup>

(1 Guangxi Climate Center, Nanning 530022; 2 Guangxi Research Institute of Meteorological Disasters Mitigation, Nanning 530022)

**Abstract:** The dimension reducing processing was conducted by means of the EOF method on a number of the environmental field and series periodic factors with definite physical meanings. In collaboration with the artificial neural network (ANN) scheme, a qualitative forecast model for the annual frequency of tropical cyclones and summer (June to August) rainfall affecting Guangxi was worked out. The results indicate that this model has a better qualitative capability of predicting the annual frequency of tropical cyclones and summer rainfall over Guangxi and can be taken as an auxiliary operational short-term climate prediction technique.

**Key words:** dimension reducing, artificial neural network, tropical cyclone, summer precipitation, qualitative forecast