

主分量逐步回归在气温预测中的应用^{*}

曾 琮¹, 陈创买²

(1 广州中心气象台, 广东 广州 510080;
2 中山大学大气科学系, 广东 广州 510275)

摘 要: 以广东省冬季气温场为预报场, 前一年的北太平洋海温场和北半球 500 hPa 高度场为因子场, 分别对它们作标准化处理, 然后进行主分量分析, 得到主分量矩阵。通过相关分析和逐步回归, 求得预报场的主分量与因子场的主分量之间的关系, 对预报场的标准化主分量进行反算, 得到原始场的拟合和预报。结果显示, 广东省冬季气温场前 4 个主分量 (对总方差的累积贡献达到 97.5%) 的预报方程都通过显著性检验, 其方程复相关系数基本在 0.9 以上。对广东省 48 个代表站 2001–2003 年冬季气温进行预测检验, 大部分预测结果的残余标准差比同期的样本标准差低, 同时回归方程的预报误差略低于实际业务预报误差, 因而回归方程对实际天气预报业务工作有一定的参考意义。

关键词: 主分量; 逐步回归; 温度; 预测

中图分类号: P458.1 **文献标识码:** A **文章编号:** 0529-6579 (2006) 04-0107-04

在气象要素场的预报方面, 不少学者进行过研究^[1-4], 主分量逐步回归是其中比较好的方法之一。主分量分析的特点就是能够将气候场的绝大部分信息集中在头几个主分量中, 因此, 通过主分量分析, 可以将气候场的预报变成对气候场主分量的预报。预报因子也是来自各因子场的主分量, 它们集中了各因子场的最重要信息, 且不同的主分量彼此是无关的。通过相关和回归分析来建立气候场与各因子场的广泛联系, 可收集到高信息量的预报因子。而气候要素场的主分量是与其空间分布型相对应的, 从而使预报量与预报因子的相关, 实际上变成气候场的某种分布型与某因子场的一定空间分布型的相关。

本文将采用海温场、高度场的主分量作为预报因子, 用气候场的主分量逐步回归方法做广东冬季气温预报, 并对广东 2001–2003 年冬季气温进行预测检验。

1 资料处理及分析原理

1.1 资料处理

1.1.1 广东气候资料。选取广东省 48 个站点 1954–2002 年共 49 a 的冬季 (12–2 月) 地面气温资料作为基本资料。由于有些站点建立较晚, 造成观测记录长度不一, 需要将这些站点资料延长到

1954 年; 还有个别站点出现缺、错记录, 需对资料进行插补。本文采用差值订正法等时间序列订正法^[5]来对资料进行延长和插补。资料取自广东省气象台。

由于广东冬季是从 12 月到 2 月, 本文是以当年 1–2 月和上一年 12 月作为本年冬季的, 如 2002 年冬季指的是 2001 年 12 月到 2002 年 2 月这段时间。

1.1.2 因子资料。因子资料包括海温场和 500 hPa 高度场的主分量, 共 1792 个因子。海温场和高度场资料取自国家气候中心。

取 1953 年 1 月到 2002 年 12 月北太平洋 286 网格点的月平均海温场资料, 将其分为 5 个区: 1 区表示太平洋西北部 (180°E 以西, 10°N 及其以北), 2 区表示太平洋东北部 (180°E 以东, 15°N 以北), 3 区表示赤道东太平洋区 (180°E 以东, 10°S~15°N), 4 区表示太平洋北部全区 (110°E 以东, 15°S 以北), 5 区表示太平洋热带区 (110°E 以东, 10°S~30°N)。将每年各月资料按春 (3–5 月)、夏 (6–8 月)、秋 (9–11 月)、冬 (12–2 月) 四季、六个自然季节 (春、初夏、盛夏、夏末、秋、冬)、前汛期 (4–6 月)、后汛期 (7–9 月)、夏季半年和冬季半年总共分为 14 个季节, 对每个区的每个季节的海温场进行主分量分析

* 收稿日期: 2005 09 28

基金项目: 广东自然科学基金资助项目 (5001121)

作者简介: 曾琮 (1970 年生), 男, 高级工程师, E-mail: zcong628@126.com

并取前 8 个主分量, 则有 560 个因子。

取 1953 年 1 月到 2002 年 12 月北半球 500 hPa 高度场梅花格点共 576 网格点的月平均高度资料, 将其分为 11 个区: 1 区表示东北半球北部 ($5^{\circ}-180^{\circ}\text{E}$ $50^{\circ}\text{N}-85^{\circ}\text{N}$), 2 区表示东北半球南部 ($5^{\circ}-180^{\circ}\text{E}$ $10^{\circ}-45^{\circ}\text{N}$), 3 区表示西北半球北部 ($180^{\circ}-0^{\circ}\text{W}$, $50^{\circ}\text{N}-85^{\circ}\text{N}$), 4 区表示西北半球南部 ($180^{\circ}-0^{\circ}\text{W}$, $10^{\circ}-45^{\circ}\text{N}$), 5 区表示东北半球 ($5^{\circ}-180^{\circ}\text{E}$ $10^{\circ}-85^{\circ}\text{N}$), 6 区表示西北半球 ($180^{\circ}-0^{\circ}\text{W}$, $10^{\circ}-85^{\circ}\text{N}$), 7 区表示北半球 ($5^{\circ}-360^{\circ}\text{E}$ $10^{\circ}-85^{\circ}\text{N}$), 8 区表示北太平洋地区南部 ($90^{\circ}-270^{\circ}\text{E}$ $10^{\circ}-45^{\circ}\text{N}$), 9 区表示北太平洋地区北部 ($90^{\circ}-270^{\circ}\text{E}$ $50^{\circ}-85^{\circ}\text{N}$), 10 区表示北半球副热带高压 ($5^{\circ}-360^{\circ}\text{E}$ $10^{\circ}-45^{\circ}\text{N}$), 11 区表示北半球极涡区 ($5^{\circ}-360^{\circ}\text{E}$ $50^{\circ}\text{N}-85^{\circ}\text{N}$)。同样取春、夏、秋、冬等 14 个季节, 并对每个区的每个季节的高度场进行主分量分析并取前 8 个主分量, 则有 1232 个因子。

1.2 分析原理

主分量分析 (PCA), 是气象上多变量分析中常用的方法之一。它能够把随时间变化的气象要素场分解为空间函数部分和时间函数 (主分量) 部分。空间函数部分概括场的地域分布特点, 这部分是不随时间变化的; 而时间函数部分则由空间点 (变量) 线性组合而成, 称为主分量, 这些主分量的头几个占有原空间点的总方差的很大一部分^[6]。由主分量的性质, 不同的主分量彼此是无关的, 可以用头几个方差较大的时间函数与其对应的空间函数乘积作为原气象要素场的估计, 也就可以将气候场的预报变成对气候场主分量的预报。

主分量逐步回归方法首先对气候场作标准化处理, 再进行 PCA 计算, 得主分量矩阵。为了提高预报效果, 预报因子也是来自各因子场的主分量。通过相关分析和逐步回归计算, 对该场的标准化主分量逐个进行拟合和预报。然后对该场的标准化主分量预报进行反算, 得到标准化变量和原始变量的拟合和预报, 并作误差估计。

1.2.1 对气候场作主分量分析

对标准化气候场 $Z_{p \times n}$ 进行 PCA 计算, 得:

$$Z_{p \times n} = V_{p \times r} W_{r \times n} \quad (1)$$

式中 V 为 $p \times r$ 阶特征向量, W 为 $r \times n$ 阶主分量, r 为 p 和 n 的最小值。

取占 95% 以上方差的前 K (≤ 15) 个主分量进行标准化, 即

$$Y_{k \times n} = \Lambda^{-\frac{1}{2}} W_{k \times n} \quad (2)$$

其中 $\Lambda^{-\frac{1}{2}}$ 为特征元素对角阵。

1.2.2 对气候场的标准化主分量进行回归预报

对预报场的前 K 个主分量逐个建立回归方程, 则第 k 个标准化主分量的回归预报模型为:

$$y_{ki} = \alpha_k + \sum_{i=1}^L C_{ki} f_{kit} + e_k, \quad t = 1, 2, \dots, n \quad (3)$$

其中 C_k 为常数, C_{ki} 为回归系数, f_{ki} 为预报因子, e_k 为误差, L 为可能预报因子数。

由于预报因子的数量很大, 在建回归方程前先要对所有预报因子进行初步相关筛选。

如果有 L_k 个预报因子进入回归方程, 则由式 (3) 得

$$\hat{y}_{ki} = \alpha_k + \sum_{i=1}^{L_k} C_{ki} f_{kit}, \quad k = 1, 2, \dots, K \quad (4)$$

1.2.3 对标准化气候场的拟合和预报

先对气候场的标准化主分量的预报值和拟合值进行反算:

$$\hat{W}_{k \times (n+s)} = \Lambda^{\frac{1}{2}} \hat{Y}_{k \times (n+s)} = \Lambda^{\frac{1}{2}} (\hat{Y}_{k \times n} \hat{Y}_{k \times s}) \quad (5)$$

根据 Gray 的研究^[7], 假定在很短的时间内, 预报场的空间函数场 $V_{p \times k}$ 是不变的, 可以对标准化变量场进行拟合和预报:

$$\begin{aligned} \hat{Z}_{p \times (n+s)} &= V_{p \times k} \hat{W}_{k \times (n+s)} = \\ V_{p \times k} (\hat{W}_{k \times n} \hat{W}_{k \times s}) &= (\hat{Z}_{p \times n} \hat{Z}_{p \times s}) \end{aligned} \quad (6)$$

1.2.4 对原始场的拟合和预报

$$\hat{X}_{p \times (n+s)} = S_p \hat{Z}_{p \times (n+s)} + E_p \quad (7)$$

其中 S 和 E 分别为原始气候场的均值和标准差序列。

2 主分量逐步回归预测广东冬季气温

用前一年北太平洋海温场和 500 hPa 高度场的主分量作为预报因子群, 通过主分量逐步回归建立预报方程预测当年广东冬季气温。为了尽可能多地利用历史资料和最新资料, 在做每年冬季气温预报时, 都根据最新资料重新建立预报方程。如预报 2001 年广东冬季气温, 用 1954 - 2000 年广东气温做样本, 因子场取 1953 - 2000 年高度场和温度场建预报方程。而做 2003 年冬季气温预报时, 则用 1954 - 2002 年广东气温做样本, 因子场取 1953 - 2002 年高度场和温度场重新建立预报方程。

2.1 广东冬季气温预测方程的建立

以下就以建立 2003 年冬季气温预报方程为例, 具体介绍方程的建立过程和结果。

2.1.1 对冬季气温场作 PCA 计算 对广东冬季气温场作 PCA 计算^[8], 发现其第 1 主分量的方差占了总方差的绝大部分, 达到 93.10%。前 4 个主分

量对总方差的累积贡献达到 97.5%, 说明用主分量分析逼近广东冬季气温场的收敛性好。显然前 4 个主分量已反映了场的主要分布特征, 因此本文选取前 4 个主分量进行分析。

2.1.2 因子初步筛选 逐个对冬季气温场的前 4 个主分量与前期预报因子进行相关计算, 适当调整显著性水平 α 的值, 使各个温度场的主分量都能从因子群中找到 65 ~ 70 个可能预报因子, 供逐步回归计算。

2.1.3 逐步回归计算 给定因子选择的 F 分布临界信度 $F_\alpha = 0.85$ 对冬季气温场的前 4 个主分量逐个进行逐步回归计算 (用初步筛选后的因子), 建立 4 个主分量的预报方程。以下仅列出第 1 主分量的回归方程。

广东冬季气温第 1 主分量的回归方程为

$$y_1 = 0.3056 + 0.7523x_1 + 0.6279x_2 + 0.5673x_3 - 0.3166x_4 + 0.4099x_5 - 0.3828x_6 + 0.4306x_7 + 0.5367x_8 + 0.3737x_9 - 0.4135x_{10} \quad (8)$$

式中, x_1 为 3 - 4 月高度场 2 区第 5 主分量; x_2 为 9 - 10 月高度场 2 区第 6 主分量; x_3 为 7 - 9 月高度场 4 区第 4 主分量; x_4 为冬季高度场 5 区第 5 主分量; x_5 为 4 - 9 月高度场 6 区第 4 主分量; x_6 为春季高度场 8 区第 2 主分量; x_7 为 4 - 9 月高度场 11 区第 8 主分量; x_8 为夏季高度场 11 区第 6 主分量; x_9 为春季海温场 4 区第 6 主分量; x_{10} 为 11 - 12 月海温场 4 区第 4 主分量。

2.1.4 拟合 利用式 (8) 对广东冬季气温第 1 主分量进行拟合, 其拟合效果如图 1 所示, 从图中可见拟合效果非常好。

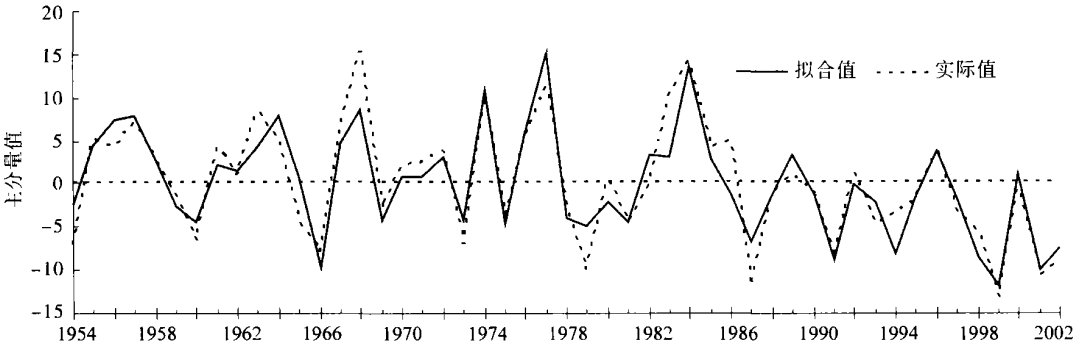


图 1 广东冬季气温第 1 主分量的拟合

Fig. 1 The simulated and observed first principle components of winter temperatures in Guangdong during 1954 - 2002

按照上述方法, 分别建立预报广东 2002 年 12 月和 2003 年 1 月、2 月气温的前 4 个主分量的回归方程。根据回归方程对 1954 - 2002 年广东冬季气温第 1 到第 4 主分量分别进行拟合, 其复相关系数基本在 0.9 以上。

2.2 广东冬季气温预测方程的检验

根据同样原理, 分别重新建立 2001 和 2002 年冬季气温预测方程。

将 2001 - 2003 年广东冬季气温场的前 4 个主分量预报结果与其对应的特征向量相配合, 可以算出各时间段 48 个站点的气温预测值, 与对应的实况值相减后可以算出各站点气温预报绝对误差值, 将各站点的绝对误差之和除以总站数可得平均绝对误差; 将绝对误差平方和除以总站数后求平方根可得到残余标准差。

对广东省 48 个站点 1954 - 2002 年的冬季 (12 - 2 月) 地面气温资料进行统计分析, 得出广东省各测站样本平均气温标准差: 冬季为 0.87 -

1.17℃, 12 月为 1.22 - 1.62℃, 1 月份为 1.23 - 1.61℃, 2 月份为 1.51 - 2.15℃。

将预测结果的残余标准差与同期的样本标准差对比, 发现有 7 个预报量的残余标准差比同期的样本标准差低, 2 个预报量在样本标准差范围附近, 还有 3 个预报量超过样本标准差, 总的预测效果还可以。其中预测最接近实况的是 2002 年冬季气温预报, 预测 2002 年广东各站冬季气温偏高 1.0 ~ 1.6℃, 实况是偏高 0.6 ~ 1.9℃, 绝大部分测站的预测值和实况值之间的误差在 0.3℃以内。

将方程算出的预测值平均绝对误差与 2001 - 2003 年实际业务发布的预报平均绝对误差对比, 结果是回归方程的预报值略优于实际业务发布的预报值, 因而具有实用价值。

3 结 语

(1) 气候场的主分量逐步回归预报模型, 将气候场的预报变成对气候场的主分量的预报。预报

因子也取自各种因子场的主分量, 可以为气候场的预报提供高信息量的因子。

(2) 预报对象和预报因子都是主分量, 两者的关系反映了气候场的某种特定分布型与因子场的某一特定分布型之间的关系, 所以模型保持着两个场相关的特点

(3) 用海温场和高度场的主分量作为预报因子, 用主分量逐步回归方法能很好地拟合广东冬季气温变化。对广东 2001—2003 年冬季气温进行试预报, 预测结果与实况较接近。

(4) 主分量逐步回归建立的预报方程对实际天气预报业务工作有一定的参考意义, 其预测效果将在今后的实际业务工作中得到进一步检验。

参考文献:

[1] 陈创买, 薛纪善, 林应河. 广东灾害性气候的分析和预

测研究[M]. 广州: 中山大学出版社, 1999

- [2] 谢炯光, 秦冰冰, 王静渊. 奇异值分解及其在季降水预报中的应用[J]. 气象学报, 1997, 55(1): 117—124
- [3] 袁景凤, 吴晓曦. 一种新的多因子综合预报方法及其预测能力评估[J]. 应用气象学报, 2000, 11(增刊): 98—102
- [4] 周文, 陈创买, 简裕庚. 广东春雨的长期预报[J]. 中山大学学报: 自然科学版, 1999, 38(4): 107—111
- [5] 么枕生, 丁裕国. 气候统计[M]. 北京: 气象出版社, 1990, 786—788
- [6] 黄嘉佑. 气象统计分析与预报方法[M]. 北京: 气象出版社, 2000, 126—147
- [7] GRAY B H. On the stability of temperature eigenvectors pattern[J]. Climatology, 1981(1): 273—281
- [8] 曾琮, 陈创买, 李晓娟. 广东冬季气温时空变化特征[J]. 气象, 2005, 31(3): 56—60

Application of Stepwise Regression Analysis of Principal Components in Temperature Prediction

ZENG Cong¹, CHEN Chuang-mai²

(1. Guangzhou Central Meteorological Observatory, Guangzhou 510080 China

2. Department of Atmospheric Science, Sun Yat-sen University, Guangzhou 510275 China)

Abstract Using the average temperature data in winter in Guangdong as the predictand field, the sea surface temperature field in the North Pacific and 500hPa geopotential heights in the Northern Hemisphere of the previous year as predictors, which are normalized, a principal component matrix is obtained by principal component analysis. The correlation between the principal components of predictand fields and the predictors is acquired by correlation and stepwise regression analyses. The fitting and forecast of primitive fields are produced by the retrieval of fields of normalized principal components. The results show that the prediction equations consisting of the first 4 principal components (97.5% of accumulated contribution to the total variance) of winter temperatures field in Guangdong pass the significance test with the multiple correlation coefficients over 0.9. Based on the forecast verification of the winter average temperature data at 48 stations in Guangdong from 2001 to 2003, it shows that most of the residual variances of the forecasting results are lower than that of the sample, and the forecast error by regression equations is lower than that of operational weather forecasting. Therefore, the results from the study have practical implication in operational weather forecasting.

Key words principal components; stepwise regression analysis; temperature; prediction