

深圳地区日极值气温的降尺度研究

武杨 李晴岚

(中国科学院深圳先进技术研究院 深圳 518055)

摘要 气象与人类日常生活的关系十分密切, 气象预报一直是人类社会高度关注的问题。随着经济的发展和社会的进步, 人类对天气预报的准确性提出了越来越高的要求, 迫切希望实现气象要素精细化预报。获取详细准确的区域气象资料是实现气象精细化预报的首要条件, 全球大气环流模式是目前预估大尺度未来全球气候变化最重要的模式, 能较好地模拟出大尺度的平均特征。但是模式预报输出的空间分辨率较低, 无法获取精细的区域气象资料, 很难对区域天气情景变化做出详细的预测, 而降尺度方法可用于弥补这方面的缺陷。文章的研究工作主要是利用统计降尺度的多元线性回归方法和BP神经网络方法对深圳地区近十年的日最低温度和最高温度进行降尺度分析研究。采用的数据是美国国家环境预报中心/美国国家大气研究中心提供的FNL全球分析资料和深圳国家基本观测站——竹子林站的实际观测数据, 重点研究了基于BP神经网络方法和多元线性回归方法的统计降尺度模型的设计与实现过程, 并对两种方法的结果进行了比较, 为区域站点的统计降尺度应用提供了设计方法和参考。

关键词 统计降尺度; 日极值气温; 多元线性回归; BP神经网络; 深圳
中图分类号 TG 156 文献标志码 A

A Downscaling Study on the Daily Temperature Extremums in Shenzhen

WU Yang LI Qinglan

(Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)

Abstract The weather has a profound influence on human's daily life and the weather forecasting has always been a topic of great concern. With the economic development and social progress, people's requirements for daily weather forecasting has become higher and higher. Information provided by the general circulation models (GCMs) can describe well some of the weather parameters at a large scale, but GCMs fail to provide detailed weather information at a regional or local scale for impact assessment studies. Outputs from GCMs are usually of low spatial resolutions. A common approach to bridge the scale mismatch is downscaling. In the present study, two methods, i.e., the statistical multiple linear regression and the BP neural network, were proposed to downscale large scale reanalysis data to daily temperature extremums at a local point, Shenzhen national meteorological station. The data used in this study are NCEP/NCAR (National Centers for Environmental Prediction/National Centre for Atmospheric Research) reanalysis dataset for the 2000-2012 period and daily observations of maximum temperature and minimum temperature at Shenzhen station for the same period. The two methods were compared in this study. Results show that both methods can simulate well the daily temperature extremums at Shenzhen station, but the performance of the statistical downscaling method is more stable than the BP neural network.

Keywords statistical downscaling; daily temperature extreme; multiple linear regression; BP neural network; Shenzhen

收稿日期: 2013-12-30

基金项目: 深圳市科技研发资金项目(JCYJ20120617115926138)。

作者简介: 武杨, 硕士研究生, 研究方向为计算机应用; 李晴岚(通讯作者), 博士, 副研究员, 研究方向为气候变化、天气预报和台风模拟与影响评估, E-mail: ql.li@siat.ac.cn。

1 引 言

随着经济的发展和社会的进步,天气对社会的发展和人们的生产生活影响越来越大。气象信息在农业生产、交通运输、资源开发、旅游商务和军事安全等各个领域中的指导作用也越来越重要。因此,社会的发展不仅要求气象预报部门能够提供传统常规的天气预报服务,而且还需要获取高分辨率的气象预报信息。降尺度方法是目前实现精细化预报、获取高分辨率气象信息的重要手段,主要可以分为基于发展更高分辨率数值预报模式的动力降尺度方法和基于已有大尺度预报产品的统计降尺度方法^[1]。这两种方法各有优缺点:动力降尺度方法物理意义较为明确,是未来气象预报领域长期发展的方向,但其计算量很大,模拟和适用范围受限制;统计降尺度方法是通过统计学手段建立大尺度和小尺度之间的尺度转换,计算量较小,经济实用,并且不受区域限制。所以,统计降尺度方法十分适合基层气象站点的推广使用,是实现精细化预报相对直接有效的手段^[2]。

Kidson 等^[3]运用了逐步回归的统计降尺度方法对新西兰地区的日最高气温和日最低气温进行了预测分析,同时也对同样的数据资料使用了耦合 RAMS 区域模式的动力降尺度方法,但两种不同的降尺度方法的结果差别不大;Murphy^[4]也使用同样的方法对所在区域进行了对比研究,得出了基本相同的结论。这说明了动力降尺度虽然耗时很大,但是在目前发展状况下并没有取得较统计降尺度具有明显优势的效果。Sailor^[5]对美国气象观测站点的气温进行了多元线性回归的统计降尺度研究;Murphy^[6]也使用了相同的方法对欧洲若干站点的气温观测值进行了模拟,其预测均取得了一定的效果;Hellstrm 等^[7]对瑞士的降水情况进行了研究,其结果相比原始的模式预报精确了很多,但是在不同季节时间存在一定的误

差;范丽军^[8]使用了逐步回归的方法对中国华北地区 49 个气象观测站的资料进行了分析,并对各个气象观测站的月平均温度进行了预测,与实际观测对比后反映预测值与实际值有比较一致的趋势,但是仍然存在部分时间段和部分气象观测点差别较大的现象。Huth 等^[9]对欧洲 8 个气象观测站秋冬季逐日温度进行统计降尺度研究,使用了多元线性回归方法并从多个角度进行评价,证明了线性方法在一般情况下也可以取得较好的降尺度效果,但是在跨月变化时会存在较大误差,Paulin^[10]使用了 SDSM (Statistical Downscaling Model) 对加拿大东南部日极端气温进行降尺度分析也存在这样的现象。

大气运动是非线性的^[11],因此,多元线性回归方法描述大气运动在理论上存在不足。为了反映大气的非线性特征,非线性的降尺度方法,如神经网络方法,也被应用到气温的降尺度分析中。

美国 Neural Ware 公司是较早开展利用神经网络进行气象预报研究的^[12],1987 年,该公司研究并设计了一套该地区基于人工神经网络的晴雨预测系统,该系统的输入是地面气压和风向资料,输出为逐日的晴雨预报,结果经过检验准确率略高于当地预报员的水平;Baik 等^[13]应用神经网络方法对热带气旋强度进行了预报,并与多元线性回归的结果进行了比较,结果表明神经网络方法效果要好于多元回归方法;美国的 David 等^[14]采用 BP 神经网络模型对加利福尼亚地区的降水进行了研究,建立加利福尼亚 7 个不同区域的预报模型,经过分析后预报结果与实际分布基本保持一致。

天气对人类日常生活影响很大,尤其是气温。气温的高低直接影响人们的感官感受。气温预报的准确性对农业的生产影响极为重要,对社会的经济发展建设、防灾减灾有着十分重要的意义。同时,气温作为最基本的气象要素,它的预

报准确性直接关系到其他气象要素和天气现象的预报准确性, 特别是灾害性天气往往和气温的变化有着密切的联系。然而, 提高气温预报的精细化程度十分困难, 目前中国的气象要素预报只能达到县地区级别的预报, 不能满足更小范围甚至某个站点的更为精细化预报需求。

基于以上背景, 本文主要利用美国国家环境预报中心(NCEP)/国家大气中心(NCAR)提供的 FNL 全球再分析资料(Final Operational Global Analysis)以及深圳市气象局实际每日温度观测数据, 应用多元线性回归以及神经网络方法构建深圳地区日气温最高值及最低值的统计降尺度模型。旨在比较这两种方法在降尺度研究中的表现, 并提高深圳地区大气数值产品的释用, 生成较高分辨率的气温预报资料, 从而改善实际预报中的预报精确度。

2 资料和方法

2.1 资料

本文构建统计降尺度模型使用的大尺度大气环流资料为 2000~2012 年美国国家环境预报中心(NCEP)/国家大气中心(NCAR)提供的 FNL 全球再分析资料。NCEP 的 FNL 数据产品格式和 NCEP 使用的全球预报系统(Global Forecast System, GFS)数据格式是一致的, 采用了表格驱动编码的 GRIB1 版本编码格式^[15]。NCEP FNL 数据的空间分辨率为 $1^\circ \times 1^\circ$, 时间间隔为 6 h(即每天 4 次, 国际时刻 00、06、12、18 时), 数据中包含了地表以及 26 个标准等压层(1000 mb~10 mb)、地表边界层(部分为 sigma 层)和对流顶层等的要素信息。NCEP 再分析资料由于充分同化了尽可能全面的观测资料, 与其他资料相比较, 是目前最为全面和可靠准确的全球大气大尺度网格数据^[16]。

本文使用的小尺度变量是深圳国家基本气象

站——竹子林站(区站号为 59493)2000 年~2012 年的实际气温观测数据。竹子林站位于北纬 $22^\circ 32'$ 、东经 $114^\circ 00'$ 、海拔 63 米, 竹子林站具有 50 多年资料积累, 观测和资料审核严格执行国家相关规范, 单点资料代表深圳参加全国气象资料交换, 资料通过中国气象局认定。

2.2 方法

统计降尺度方法主要依赖于小尺度区域数年的实际气象要素的观测资料。通过统计学原理建立实际观测的气象要素与大尺度大气环流因子之间的关系, 然后把全球大气环流模式模拟的大尺度预测环流因子信息作为这种关系的输入, 得到小尺度区域气象要素变量, 例如气温等的变化趋势^[17]。

统计降尺度的关键在于利用统计学方法建立大尺度大气环流预报因子和区域气象要素预报变量之间的映射关系 $Y=F(X)$, 其中 X 表示大尺度大气环流预报因子, Y 表示区域气象要素预报变量。 F 则表示了从多年区域的实际观测资料中提取出来的表达区域地形、下垫面等区域独特的地理环境特征, 以及小尺度区域受到大尺度网格的大气环流影响的特征等复杂信息。统计降尺度法的优点主要是能够纠正大气环流模式(General Circulation Model, GCM)的系统误差, 根据区域气候信息的特征将全球大气环流模式中相关性较大, 物理意义较明显, 预测较为准确的大气环流因子筛选出来用于统计模型中, 因而不必考虑边界初值条件的影响^[8]。本文将采用多元线性回归的方法和 BP 神经网络的方法来建立映射关系。

2.2.1 多元线性回归方法

假设 y 是深圳地区日气温值, 表示可观测的随机变量, 它受到 p 个大尺度大气环流因子 x_1, x_2, \dots, x_p 和随机因素 ε 的影响, 若 y 与 x_1, x_2, \dots, x_p 有如下线性关系:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (1)$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 是 $p+1$ 个未知参数, ε 是不可测

的随机误差, 且 $\varepsilon \sim N(0, \sigma^2)$ 。式(1)通常称为多元线性回归模型。

根据多年的实际气温观测数据和对应的 NCEP FNL 资料, 得到 n 组样本数据 $(x_{i1}, x_{i2}, \dots, x_{ip}; y_i)$, $i=1, 2, \dots, n$, 他们满足式(1), 即有

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ \dots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{cases} \quad (2)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 相互独立且都服从 $N(0, \sigma^2)$ 。采用最小二乘法来估计多元线性回归方程中的未知参数 $\beta_0, \beta_1, \dots, \beta_p$, 即选择 $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ 使误差平方和

$$Q(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2 \quad (3)$$

达到最小。从而可以得到多元线性回归方程^[18,19]。

2.2.2 BP 神经网络方法

相对于多元线性回归方法, BP 神经网络模型属于一种高度非线性映射, 目前证明一个三层神经网络可以实现任一连续函数或映射。BP 神经网络模型结构见图 1^[20]。

BP 算法的过程如下^[21,22]:

(1) 通过网络将输入向前传播:

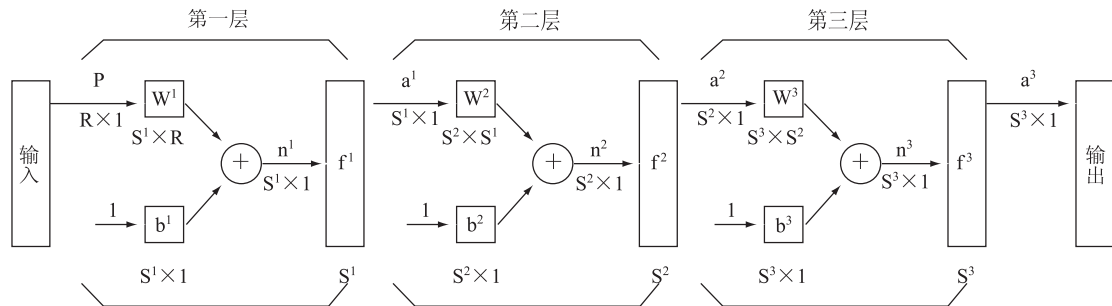


图 1 BP 网络模型结构图

Fig.1. Structure of BP neural network model

$$\mathbf{a}^0 = \mathbf{p} \quad (4)$$

$$\mathbf{a}^{m+1} = \mathbf{f}^{m+1}(\mathbf{W}^{m+1} \mathbf{a}^m + \mathbf{b}^{m+1}), m=0, 1, \dots, M-1 \quad (5)$$

$$\mathbf{a} = \mathbf{a}^M \quad (6)$$

(2) 将敏感性反向传播

$$\mathbf{s}^M = -2\dot{\mathbf{F}}^M(\mathbf{n}^M)(\mathbf{t} - \mathbf{a}) \quad (7)$$

$$\mathbf{s}^m = \dot{\mathbf{F}}^m(\mathbf{n}^m)(\mathbf{W}^{m+1})^T \mathbf{s}^{m+1}, m=M-1, \dots, 2, 1 \quad (8)$$

(3) 使用最速下降法更新权值和偏置值^[16]

$$\mathbf{w}^m(k+1) = \mathbf{w}^m(k) - \alpha \mathbf{s}^m(\mathbf{a}^{m-1})^T \quad (9)$$

$$\mathbf{b}^m(k+1) = \mathbf{b}^m(k) - \alpha \mathbf{s}^m$$

(4) 考虑到最速下降法更新权值和偏置值的速度很慢且容易陷入最小局部点, 本文采用牛顿法的速度和保证收敛的最速下降法的一个折中, Levenberg-Marquardt 算法, 采用^[23]

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{J}^T(\mathbf{x}_k)\mathbf{J}(\mathbf{x}_k) + \mu_k \mathbf{I}]^{-1} \mathbf{J}^T(\mathbf{x}_k)\mathbf{v}(\mathbf{x}_k) \quad (10)$$

或

$$\Delta \mathbf{x}_k = -[\mathbf{J}^T(\mathbf{x}_k)\mathbf{J}(\mathbf{x}_k) + \mu_k \mathbf{I}]^{-1} \mathbf{J}^T(\mathbf{x}_k)\mathbf{v}(\mathbf{x}_k)$$

来改进 BP 神经网络^[24,25]。

3 统计降尺度模型的建立

3.1 模型的建立流程

统计降尺度模型的构建首先建立全球大气大尺度网格点的环流因子和地面观测站气象要素变量之间的降尺度统计关系, 然后利用模式输出的格点预测数据和已经建立的统计关系, 得到未来

情景下地面测站的变量预测值。整个降尺度方法的过程为:

(1) 选择预报变量

根据降尺度的目的, 确定所要降尺度的地理区域和预报变量。收集该地理区域实际的观测资料, 并且分析预报变量资料的质量状况, 修补和订正数据的缺漏和错误。本文选择深圳市国家基本观测站——竹子林站的实际观测数据, 分析和预测的气象变量是日气温的最高值和最低值。

(2) 选择预测因子

这是降尺度过程中最为重要的步骤之一, 因为选择的预测因子的好坏直接关系到模型的成功与否。因此, 本文通过分析影响气温变化的物理原理和运用相关分析的方法去筛选大气环流因子作为降尺度的预测因子。

(3) 降尺度的方法选择

统计降尺度的方法有很多, 不同的方法所取得的效果也是不同的, 尤其是针对不同的预报变量。本文将采用逐步多元回归的方法和 BP 神经网络的方法进行研究和比较。

(4) 模型的建立和检验

通过训练样本的数据输入到降尺度选择的方法中, 建立起降尺度模型。对模型的检验包括原有训练样本数据的输入拟合状况检验及验证样本数据的拟合状况, 其中最主要的是对验证样本数据的检验, 这是考察建立的模型泛化能力是否良好的关键。如果模型的泛化能力不好就要考虑重新选择降尺度的方法或者重新建立模型。

(5) 应用模型

将 GCM 输出的未来大气环流因子应用到建立好的模型中, 生成预报变量的预测结果, 然后再进行结果的分析讨论。

3.2 预报变量的划分

模型的训练样本与检测样本属于同分布数据是统计降尺度分析的前提条件。只有训练样本和检测样本属于同分布的数据, 模型通过训练样本

学习到的内在规律才值得推广, 才会具有较好的泛化能力和预测能力。假如训练样本与预测样本不属于同一分布, 这两个样本的内在规律本身就不同, 通过训练样本学习到的规律与预测样本本身的规律也就不相同, 应用不相同的规律去做预测其结果必然差别很大。

因此, 本文采用“分而治之”的思想对研究对象按照分布特征的规律不同划分不同的集合分别进行研究。气象学上普遍认为, 某一区域多年同月份的气候状况是具有相似性的。考虑到日最高温度和最低温度也属于不同的类别, 本文将采用按月份建立日最高温度和最低温度统计降尺度模型的划分方法。这样的划分可以减小研究对象内在规律的复杂程度, 使得模型只学习属于同月份分布的样本, 学习目的明确, 不会受到其他不同月份分布样本规律的影响。

将 2000~2012 年共 13 年的数据按月份和日最高最低温度的不同划分为 24 个数据子集合, 即深圳地区 2000~2012 年 1 月份日最高气温数据集, 2000~2012 年 1 月份日最低气温数据集, 2000~2012 年 2 月份日最高气温数据集, 2000~2012 年 2 月份日最低气温数据集, 依次类推至 2000~2012 年 12 月份日最高气温数据集, 2000~2012 年 12 月份日最低气温数据集等共 24 个集合。从而达到了对原有大规模数据样本“分而治之”的目的。划分为较小的数据集合之后每个集合还需要划分为训练集合和测试集合。这 24 个数据集合中每个集合都包含了 13 年的气温数据, 考虑到模型的建立需要足够多的训练样本才能够较好地掌握其中的规律, 因此采用 2000~2009 年共 10 年的数据作为训练样本集去训练网络, 2010~2012 年共 3 年的数据则作为测试样本集合来验证神经网络的学习情况。

3.3 预报因子的筛选

适当的预报因子是统计预报好坏的关键因素。虽然影响预报对象的因素很多, 但不同因素

对预报对象的影响程度大小亦相差很大，因此我们必须分析出其中的主要因素，筛选出高质量的预报因子，来获取足够多的预报信息。

某种气象现象的产生、气象要素的变化是由其物理机制所决定的，因此首先应当考虑从预报对象产生的物理机制及变化的物理过程中寻找预报因子^[26]。

局地温度变化的物理过程可以用热流量方程表示：

$$\frac{\partial T}{\partial t} = -V \cdot \nabla T - \omega(\gamma_d - \gamma) + \frac{\gamma_d}{\rho g} \left(\frac{\partial P}{\partial t} + V \cdot \nabla p \right) + \frac{1}{c} \frac{d\bar{Q}}{dt} \quad (11)$$

式中 $\frac{\gamma_d}{\rho g} \left(\frac{\partial P}{\partial t} + V \cdot \nabla p \right)$ 项，因为变压和气压平流引起的温度局地变化很小，在实际中可以不必考虑。其他三项分别为：温度平流项 $-V \cdot \nabla T$ ；垂直运动项 $-\omega(\gamma_d - \gamma)$ 和非绝热因子 $\frac{1}{c} \frac{d\bar{Q}}{dt}$ 。对于定点定站的区域温度预报而言，主要考虑大气环流状况和下垫面性质对它的影响，其主要影响因素有：低空相对湿度、降水、风、大气层结稳定性和下垫面等。

通过以上的分析，我们根据 F 达到 0.05 相关显著水平的标准以及影响气温变化物理过程的因素，从 NCEP FNL 资料提供的 137 个大气环流因子中选择出 23 个与气温最相关的因子作为备选因子，具体如表 1 所示。

这 23 个备选因子是与预报量有一定物理联系且存在着相关的备选因子。这些因子对预报量都有一定的方差贡献，但是，这并不意味着输入模型的因子越多越好。而备选因子之间存在或多或少交叉关系，因此它们对预报量的贡献多少也存在着交叉现象。并且考虑到直接将 23 个变量输入到降尺度模型中仍然存在太多变量，相关性不好，冗余信息过多等问题。需要按月份和最

表 1 NCEP 数据输出变量列表

Table. 1 List of NCEP data variables

NO	预测因子
1	PRESmsl 平均海平面压力[Pa]
2	SPFH2m 地面 2 m 比湿[kg/kg]
3	Surface-UGRDprs 表面 u 分量风[m/s]
4	Surface-VGRDprs 表面 v 分量风 [m/s]
5	Surface-ABSVprs 表面绝对涡度[s]
6	Surface-RHprs 表面相对湿度[%]
7	Surface-HGTprs 表面位势高度 [gpm]
8	Surface-hdivg(u,v) 表面散度
9	Surface-mag(u,v) 表面风强
10	500 hPa-UGRDprs 500 hPa 层 u 分量风[m/s]
11	500 hPa-VGRDprs 500 hPa 层 v 分量风[m/s]
12	500 hPa-ABSVprs 500 hPa 层绝对涡度[s]
13	500 hPa-RHprs 500 hPa 层相对湿度[%]
14	500 hPa-HGTprs 500 hPa 层位势高度 [gpm]
15	500 hPa-hdivg(u,v) 500 hPa 层散度
16	500 hPa-mag(u,v) 500 hPa 层风强
17	850 hPa-UGRDprs 850 hPa 层 u 分量风[m/s]
18	850 hPa-VGRDprs 850 hPa 层 v 分量风[m/s]
19	850 hPa-ABSVprs 850 hPa 层绝对涡度[s]
20	850 hPa-RHprs 850 hPa 层相对湿度[%]
21	850 hPa-HGTprs 850 hPa 层位势高度[gpm]
22	850 hPa-hdivg(u,v) 850 hPa 层散度
23	850 hPa-mag(u,v) 850 hPa 层风强

低最高温度的不同，进一步将这 23 个变量进行筛选，寻求最优合适的因子。

逐步回归方法是最简便和传统的筛选方法。下面给出使用逐步回归方法从 23 个备选因子中筛选出对应 12 个月的最高温度和最低温度模型的最显著相关因子。逐步回归通过引入所有 23

表 3 深圳各月最低气温变量选择表

Table 3. Significant predictors for daily T_{\min} in 12 months

	各月选择因子变量 T_{\min}											
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
比湿	√	√	√	√	√	√	√	√	√	√	√	√
地表面压力										√		
500 hpa 风强	√	√										
850 hpa 风强	√	√	√		√	√	√	√	√	√	√	√
Surface 风强							√	√	√			
500 hpa 风速 u			√	√								
850 hpa 风速 u			√									
Surface 风速 u		√								√		
500 hpa 风速 v												
850 hpa 风速 v	√		√	√					√			
Surface 风速 v		√					√			√		
绝对涡度 500 hpa						√		√				√
绝对涡度 850 hpa												√
绝对涡度 Surface				√		√						
散度 500 hpa										√		
散度 850 hpa		√									√	
散度 Surface	√			√								√
位势高度 500 hpa					√	√	√					
位势高度 850 hpa	√	√	√	√	√				√	√	√	√
位势高度 Surface	√	√	√	√	√				√	√	√	√
相对湿度 500 hpa						√	√	√	√			√
相对湿度 850 hpa				√	√	√		√	√			√
相对湿度 Surface	√	√	√	√	√	√	√	√	√	√	√	√

是影响每个月份最高温度和最低温度最重要的相关因子;

(2)除了(1)中的两个因子外,影响12个月的最高温度最显著相关因子是相对湿度850 hpa;影响12个月最低温度最显著相关因子是

850 hpa 风强;

(3)影响7月、8月、9月最低温度和最高温度最显著相关因子还有位势高度500 hpa,相对湿度500 hpa;

(4)12个月中,除了7月、8月和9月外,

其他月份最低温度和最高温度最显著影响相关因子还有位势高度 850 hpa, 位势高度 Surface。

3.4 统计降尺度模型的建立

本节建立的统计降尺度模型是基于 3.1 节的流程分析和 3.2 节、3.3 节对预报变量的划分和预报因子的筛选结果。采用的降尺度方法主要是 BP 神经网络方法和逐步多元线性回归方法。

3.4.1 逐步多元回归方法

逐步多元回归方法根据表 2 和表 3 的筛选因子结果, 分别建立最高温度和最低温度对应的多元线性回归方程, 模型的参数由最小二乘方法解方程得到。此处列出 1 月和 7 月的最低温度、最高温度的回归方程作为例子。

$$Y_{1\min} = 0.357 + 0.432 \times X_1 - 0.0881 \times X_3 + 0.224 \times X_4 - 0.0578 \times X_{10} - 0.167 \times X_{17} + 1.004 \times X_{19} - 1.416 \times X_{20} - 0.251 \times X_{23}$$

$$Y_{7\min} = 0.204 + 0.484 \times X_1 + 0.246 \times X_4 + 0.122 \times X_5 - 0.329 \times X_{11} + 0.535 \times X_{18} - 0.153 \times X_{21} - 0.398 \times X_{23}$$

$$Y_{1\max} = 0.288 + 0.347 \times X_1 - 0.0960 \times X_3 - 0.0803 \times X_4 - 0.127 \times X_9 - 0.0649 \times X_{12} + 0.943 \times X_{19} - 1.161 \times X_{20} - 0.146 \times X_{22} - 0.206 \times X_{23}$$

$$Y_{7\max} = 0.408 + 0.388 \times X_1 + 0.345 \times X_2 + 0.200 \times X_5 - 0.368 \times X_8 - 0.291 \times X_{11} + 0.409 \times X_{18} - 0.180 \times X_{21} - 0.158 \times X_{22} - 0.410 \times X_{23}$$

3.4.2 BP 神经网络方法

同时, 为了研究非线性方法在该领域的应用效果, 采用逐步回归筛选因子组合 BP 神经网络的方法^[28]建立新的统计降尺度模型。

(1) 归一化处理

该模型将逐步回归筛选出来的预测因子作为神经网络的输入变量, 考虑到每种大气环流因子都具有不同的物理意义和不同的量纲, 比如绝对涡度分量在 $10^{-6} \sim 10^{-4}$ 范围内变化, 地表面压力分量则在 $10^4 \sim 10^6$ 范围内变化。如此大的数量级差别直接输入网络会使网络在学习过程中权值的

变化不均匀, 容易出现网络的不稳定和提前进入训练结束状态。必须通过变换处理将网络的输入输出数据限制在 $[-1, 1]$ 区间内, 使每个输入分量对网络的影响处在相同重要的地位, 从而避免权值变化不均匀的情况出现。

因此采用如下区间变换式将输入-输出数据变换为 $[-1, 1]$:

$$y = (y_{\max} - y_{\min}) * \frac{(x - x_{\min})}{(x_{\max} - x_{\min})} + y_{\min} \quad (12)$$

其中

$$y_{\min} = -1, y_{\max} = 1 \quad (13)$$

按上述的方法变换后, x 序列的最大值变为 1, 最小值变为 -1。

(2) 隐含层节点的设计

理论分析证明, 三层 BP 神经网络足以映射所有连续的函数, 只有当需要映射的函数很复杂或者不连续的时候才需要两个及以上的隐含层, 所以, 为了避免网络过于复杂, 本文考虑设置一个隐含层。

对于隐含层节点数的确定问题, 至今仍无法用一个解析式精确求得, 因此, 1990 年, Eberhart 等^[29]总结说“隐节点的选择是一门艺术”是非常准确的。隐含层节点的作用是存储从训练样本中不断学习到的内在规律, 隐含层节点的权值能够增强网络映射的能力。如果加入隐含层节点较少, 那么网络比较简单, 对于较为复杂的问题就很难从样本的学习过程中掌握足够的信息, 学习到其中的规律; 但若隐含层节点的数量过多, 网络将变得很复杂, 泛化能力反倒会降低, 网络在学习过程中对于训练样本的全部内容包括非主要的规律甚至噪声的干扰也记住, 在对训练样本外的数据进行仿真时效果并不好。

因此, 设置多少个隐含层节点是比较复杂的问题, 也是神经网络设计的困难所在之处, 主要影响的因素有映射的问题复杂程度、训练样本的

表4 深圳各月最优隐含层神经网络节点数列表

Table 4. Optimal number of the hidden layer of neural network for 12 months

月份	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
最高温度(°C)	10	12	11	10	10	12	12	11	11	13	10	11
最低温度(°C)	12	11	12	13	9	11	10	12	12	11	9	12

划分情况以及输入输出个数等。一般来说, 需要描述的问题越复杂, 规模越庞大, 网络需要的隐含层节点越多, 甚至需要增加隐含层的数目来增强其映射能力。

目前, 确定最佳隐含节点数可以根据一些经验公式作为参考, 下面的公式是一些计算隐含层节点数的经验公式:

$$m = \sqrt{n+l} + a; \quad m = \log_2^n \quad (14)$$

$$m = \sqrt{nl}; \quad \sum_{i=0}^n C_m^i > k$$

其中 k 为样本数, 如果 $i > n_1$, $C_m^i = 0$ 式中 m 为隐含层节点, n 为输入层节点数, l 为输出层节点数, a 为 1~10 之间的常数。本节根据经验公式算出隐含层节点个数大致范围, 主要采用试凑的办法, 通过设置较小的隐含层节点, 然后逐步增加个数, 通过同一样本集的反复训练得出表现最优的节点个数。表 4 所示为各个月份最低温度和最高温度模型的最优节点个数。

(3) 网络结构的设计

通过以上的分析, 以 1 月份最低气温模型对应的网络设计为例, 网络采用输入层为 8 个神经元, 中间层经过黄金分割点算法的计算确定最优节点数为 10 个神经元, 输出层为 1 个神经元。因此, 网络应该为 $8 \times 10 \times 1$ 的结构, 训练的算法采用 Levenberg-Marquardt 算法, 中间层神经元传递函数采用双曲正切 S 型传递函数 tansig , 输出层神经元传递函数采用线性传递函数, 前 10 年的数据作为训练样本, 后 3 年的数据作为检验

样本。本文采用 matlab 平台来实现神经网络的构建, 设计好的网络结构如下图 2 所示。

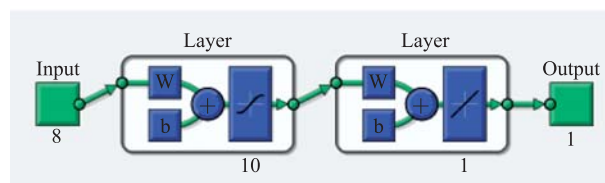


图2 BP神经网络结构图

Fig. 2. Structure of BP neural network

4 统计降尺度模型的结果分析评价

无论是神经网络模型还是多元线性回归模型, 对模型的评价主要分为两部分, 第一部分: 对训练样本集的模拟状况, 可以检验模型对训练样本的学习状况, 如果训练样本拟合得较好说明模型对于训练样本作为先验经验, 学习掌握了隐含在其中的主要规律和知识; 第二部分: 对测试样本集的预测状况, 这是检验预测模型是否具有泛化能力的关键。因为测试样本是未被模型学习过的数据集, 这部分数据集与训练样本具有同分布状态, 只有具有泛化能力的模型才能对这部分数据做出较好的预测。因此, 检验降尺度模型的好坏关键是对测试样本集预测状况的评价分析。本节采用同一评价标准对多元线性回归模型和神经网络模型结果进行评价分析。

采用的评判系数为:

(1) 相关系数平方:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \quad (15)$$

该参数是相关系数 R^2 描述了预测值和实际值的拟合程度。 R^2 被定义为回归平方和 (Regression Sum of Squares, SSR) 与总平方和 (Total Sum of Squares, SST) 的比值。 R^2 可以取值在 $[0,1]$ 之间, 越接近 1 越好, 比如 $R^2=0.85$ 表示预测数据的效果, 平均解释了原始数据 85% 的方差状况。

(2) 平均绝对误差:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (16)$$

(3) 均方根误差:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{v}} \quad (17)$$

记神经网络模型为模 A, 多元线性回归模型为模 B, 这两种模型对训练样本的拟合情况通过参数比较如表 5 和表 6 所示。

检验样本集的检验参数比较如表 7 和表 8 所示。

表 5 深圳最低气温两种模型训练样本参数比较

Table 5. Performance comparison between two models for calibrated simulation of daily T_{\min}

T_{\min}	R^2		RMSE		MAE	
	模 A	模 B	模 A	模 B	模 A	模 B
1 月	0.96	0.89	0.61	1.13	0.46	0.88
2 月	0.97	0.89	0.51	1.19	0.37	0.96
3 月	0.95	0.88	0.68	1.09	0.50	0.88
4 月	0.96	0.87	0.51	1.02	0.36	0.82
5 月	0.91	0.76	0.52	0.89	0.39	0.70
6 月	0.90	0.70	0.44	0.80	0.33	0.63
7 月	0.85	0.56	0.46	0.79	0.33	0.60
8 月	0.84	0.53	0.48	0.84	0.35	0.65
9 月	0.91	0.67	0.42	0.84	0.29	0.66
10 月	0.96	0.79	0.37	0.83	0.27	0.65
11 月	0.96	0.89	0.59	1.02	0.44	0.82
12 月	0.97	0.87	0.48	1.15	0.34	0.92

表 6 深圳最高气温两种模型训练样本参数比较

Table 6. Performance comparison between two models for calibrated simulation of daily T_{\max}

T_{\max}	R^2		RMSE		MAE	
	模 A	模 B	模 A	模 B	模 A	模 B
1 月	0.95	0.88	0.91	1.49	0.60	1.14
2 月	0.95	0.88	1.00	1.70	0.70	1.34
3 月	0.95	0.86	0.86	1.50	0.62	1.18
4 月	0.93	0.85	0.99	1.42	0.75	1.14
5 月	0.92	0.72	0.71	1.36	0.50	1.06
6 月	0.81	0.64	1.20	1.60	0.74	1.25
7 月	0.90	0.67	0.77	1.45	0.56	1.14
8 月	0.86	0.65	0.95	1.50	0.69	1.15
9 月	0.76	0.67	1.20	1.41	0.89	1.07
10 月	0.94	0.79	0.53	1.04	0.35	0.80
11 月	0.97	0.87	0.56	1.25	0.37	0.97
12 月	0.96	0.88	0.72	1.33	0.51	1.03

表 7 深圳最低气温两种模型检验样本参数比较

Table 7. Performance comparison between two models for validated simulation of daily T_{\min}

T_{\min}	R^2		RMSE		MAE	
	模 A	模 B	模 A	模 B	模 A	模 B
1 月	0.83	0.90	1.51	1.08	1.23	0.85
2 月	0.85	0.93	1.77	1.35	1.40	1.05
3 月	0.88	0.88	1.38	1.35	1.09	1.05
4 月	0.72	0.85	1.56	1.09	1.22	0.87
5 月	0.61	0.82	1.34	0.88	1.06	0.69
6 月	0.33	0.64	1.61	0.99	1.17	0.77
7 月	0.37	0.60	1.16	0.80	0.84	0.65
8 月	0.15	0.33	1.09	0.89	0.86	0.70
9 月	0.35	0.57	1.32	1.01	1.01	0.73
10 月	0.55	0.85	1.87	1.14	1.48	0.90
11 月	0.80	0.89	1.62	1.22	1.27	0.95
12 月	0.68	0.87	2.19	1.38	1.68	1.07

表 8 深圳最高气温两种模型检验样本参数比较

Table 8. Performance comparison between two models for validated simulation of daily T_{\max}

T_{\max}	R2		RMSE		MAE	
	模 A	模 B	模 A	模 B	模 A	模 B
1 月	0.74	0.80	2.24	1.87	1.83	1.55
2 月	0.81	0.90	2.48	1.68	1.99	1.38
3 月	0.72	0.89	2.47	1.46	1.98	1.15
4 月	0.75	0.80	1.83	1.64	1.38	1.24
5 月	0.43	0.52	2.13	1.65	1.58	1.29
6 月	0.32	0.70	2.34	1.39	1.83	1.11
7 月	0.35	0.57	1.95	1.47	1.57	1.14
8 月	0.19	0.20	2.14	1.97	1.60	1.43
9 月	0.54	0.64	1.64	1.48	1.24	1.13
10 月	0.71	0.79	1.59	1.29	1.23	1.02
11 月	0.73	0.86	1.78	1.17	1.28	0.92
12 月	0.80	0.88	1.88	1.52	1.47	1.17

2012 年的具体预测结果如图 3、图 4 所示。

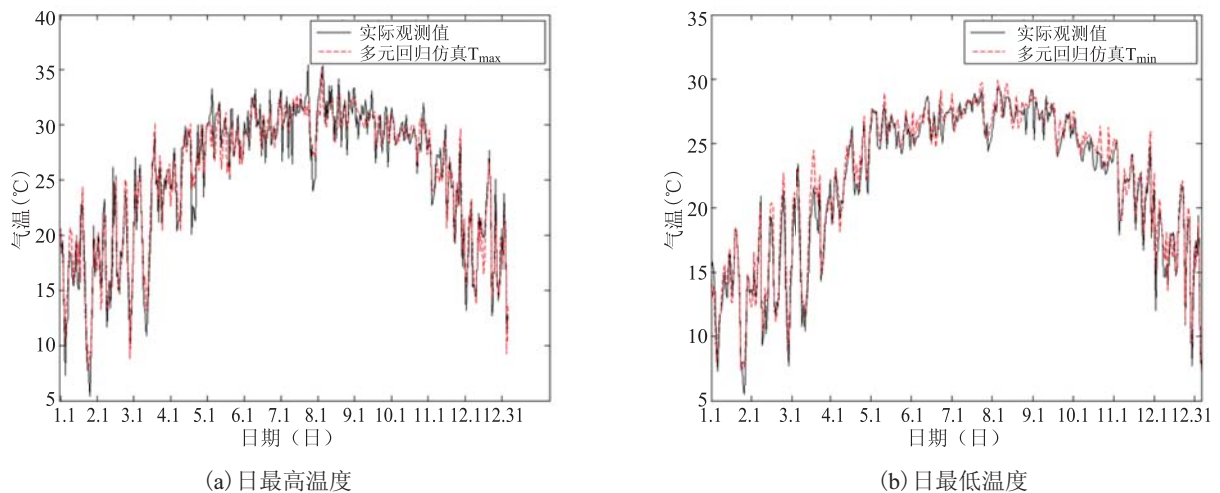


图 3 2012 年日极值温度多元回归模拟与观测比较图

Fig. 3. Comparison between the observed daily temperature extremes in 2012 and the simulated temperature by the multiple linear regression model

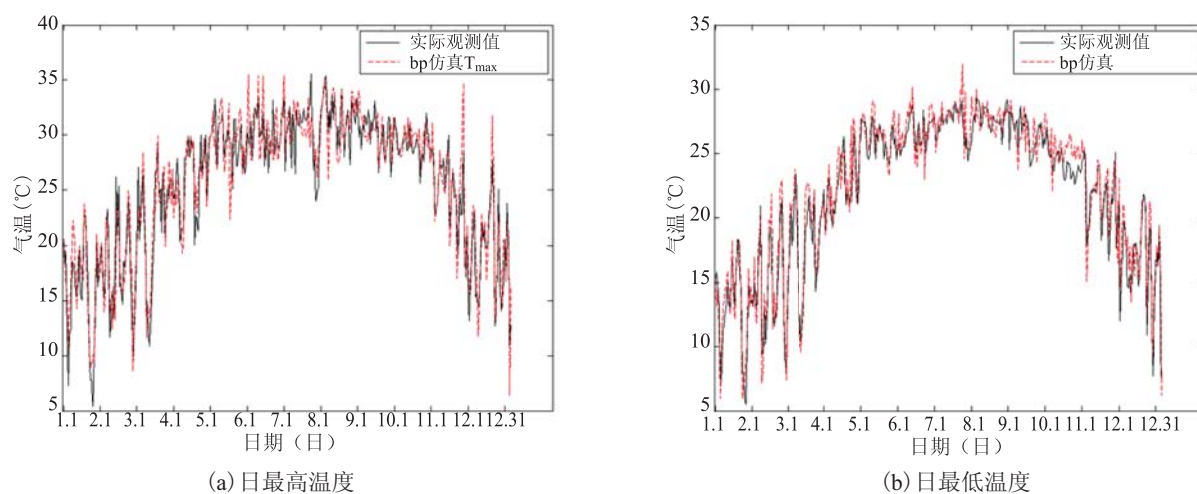


图4 2012年日极值温度BP神经网络模拟与观测比较图

Fig. 4. Comparison between the observed daily temperature extremums in 2012 and the simulated temperature by BP neural network model

通过以上数据表格和图的分析, 可以得出以下结论:

(1) 神经网络模型的训练样本拟合情况要好于多元线性回归, 但是检验样本的拟合情况却相反。

这说明了在统计降尺度领域, 神经网络并不具有比多元线性回归更好的优势, 尽管神经网络在训练样本拟合的相当好, 但是, 神经网络的过拟合问题却比多元回归突出很多, 其非线性拟合能力并没有很好的发挥出来。

(2) 两种模型对最低温度和拟合效果要好于对最高温度的拟合效果, 可以得出最高温度受到的高空环流因子影响因素较最低温度受到的影响更为复杂。

(3) 两种模型普遍对于冬季的气温模拟的效果好于夏季的气温, 主要是由于夏季深圳地区多受台风的影响, 天气变化很剧烈, 因此模型对于更为复杂的问题学习效果有待提高, 需要对数据做进一步的处理或者引入台风影响等其他相关因子。

参考文献

- [1] Dibike YB, Coulibaly P. Validation of hydro-logic models for climate scenario simulation: the case of Saguenay watershed in Quebec [J]. *Hydrological Processes*, 2007, 21(23): 3123-3235.
- [2] Wilby RL, Charles SP, Zorita E, et al. Guidelines for use of climate scenarios developed from statistical downscaling methods [OL]. http://scholar.google.ca/scholar?cluster=8727505470784486273&hl=zh-CN&as_sdt=0,5, 2004.
- [3] Kidson JW, Thompson CS. A comparison of statistical and model-based downscaling techniques for estimating local climate variations [J]. *Journal of Climate*, 1998, 11(4): 735-753.
- [4] Murphy J. An evaluation of statistical and dynamical techniques for downscaling local climate [J]. *Journal of Climate*, 1999, 12(8): 1156-2284.
- [5] Sailor DJ, Li XS. A semiempirical downscaling approach for predicting regional temperature impacts associated with climatic change [J]. *Journal of Climate*, 1999, 12(1): 103-114.
- [6] Murphy J. Prediction of climate change over

- Europe using statistical and dynamical downscaling techniques [J]. *International Journal of Climatology*, 2000, 20 (5): 489-501.
- [7] Hellström C, Chen DL, Achberger C, et al. Comparison of climate change scenarios for Sweden based on statistical and dynamical downscaling of monthly precipitation [J]. *Climate Research*, 2001, 19(1): 45-55.
- [8] 范丽军, 符淙斌, 陈德亮. 统计降尺度法对华北地区未来区域气温变化情景的预估 [J]. *大气科学*, 2007, 131(15): 887-897.
- [9] Huth R. Statistical downscaling in central Europe: evaluation of methods and potential predictor [J]. *Climate Research*, 1999, 13: 91-101.
- [10] Paulin C. Downscaling daily extreme temperatures with genetic programming [J]. *Geophysical Research Letters*, 2004, 31(16): L16203.
- [11] Hakami A, Odman MT, Russell AG. Nonlinearity in atmospheric response: a direct sensitivity analysis approach [J]. *Journal of Geophysical Research: Atmospheres*, 2004, 109(D15), doi: 10.1029/2003JD004502.
- [12] 金龙. 神经网络气象预报建模理论方法和应用 [M]. 北京: 气象出版社, 2005.
- [13] Baik LJ, Hwang HS. Tropical cyclone intensity prediction using regression method and neural network [J]. *Journal of the Meteorology Society of Japan*, 1998, 76(5): 711-717.
- [14] Silverman D, Dracup JA. Artificial neural network and long range precipitation in California [J]. *Journal of Applied Meteorology*, 2000, 39(1): 57-66.
- [15] 孙修贵. 表格驱动码编码手册—BUFR GRIB 和 CREX编码 [M]. 北京: 气象出版社, 2010.
- [16] Kalney E, Kanamitsu M, Kistler R, et al. The NCEP/NCAR 40-year reanalysis project [J]. *Bulletin of the America Meteorological Society*, 1996: 437-471.
- [17] Benestad RE, Inger HB, Chen DL. *Empirical-Statistical Downscaling* [M]. World Scientific, 2008.
- [18] Kleinbaum DG, Kupper LL, Muller KE, et al. *Applied Regression Analysis and Other Multivariable Methods* (third edition) [M]. 北京: 机械工业出版社, 2003, 250-251.
- [19] Montgomery DC, Peck EA. *Introduction to Linear Regression Analysis* [M]. John Wiley&Sons, 1982.
- [20] Hagan MT, Demuth HB, Beale MH. *Neural Network Design* [M]. Stillwater: Martin Hagan, 1996.
- [21] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors [J]. *Nature*, 1986, 323: 533-536.
- [22] Rumelhart DE, McClelland JL. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* [M]. Cambridge: MIT Press, 1986.
- [23] Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm [J]. *IEEE Transactions on Neural Networks*, 1994, 5(6): 989-993.
- [24] Scales LE. *Introduction to Non-linear Optimization* [M]. New York: Springer-Verlag, 1985.
- [25] Miller WT, Sutton RS, Werbos PJ. *Neural Networks for Control* [M]. Cambridge: MIT Press, 1990: 171-178.
- [26] 朱乾根, 林锦瑞, 寿绍文, 等. *天气学原理和方法* (第四版) [M]. 北京: 气象出版社, 2007.
- [27] Lomax RG, Hahs-Vaughn DL. *Statistical Concepts: a Second Course* (third edition) [M]. Routledge Academic Press, 2007.
- [28] Zhang DD. *Neural Networks System Design Methodology* [M]. Beijing: TsingHua University Press, 1996.
- [29] Eberhart EC. *Neural Network PC Tools: a Practical Guide* Russell [M]. New York: Academic Press, 1990.